



**HYBRIDISED INDEXING FOR RESEARCH BASED INFORMATION RETRIEVAL**

by

**KYLE ANDREW FITZGERALD**

**Thesis submitted in fulfilment of the requirements for the degree**

**Doctor of Information and Communication Technology: Information  
Technology**

**in the Faculty of Informatics and Design**

**at the Cape Peninsula University of Technology**

**Supervisor:** Prof AC de la Harpe

**Co-supervisor:** Prof AJ Bytheway

**Co-supervisor:** Dr CS Uys

**Cape Town**

**November 2019**

**CPUT copyright information**

The thesis may not be published either in part (in scholarly, scientific or technical journals), or as a whole (as a monograph), unless permission has been obtained from the University.

## DECLARATION

I, Kyle Andrew Fitzgerald, declare that the contents of this thesis represent my own unaided work, and that the thesis has not previously been submitted for academic examination towards any qualification. Furthermore, it represents my own opinions and not necessarily those of the Cape Peninsula University of Technology.



06/11/2019

---

**Signed**

---

**Date**

## EDITOR'S CERTIFICATE

5 November 2019

**KYLE ANDREW FITZGERALD**

Faculty of Faculty of Informatics and Design  
Cape Peninsula University of Technology  
Cape Town

**RE: CERTIFICATE - TECHNICAL EDITING AND PROOFREADING OF DOCTORAL THESIS**

I, the undersigned, herewith certify that the technical editing and proofreading of the D.ICT thesis of Kyle Andre Fitzgerald, entitled "*HYBRIDISED INDEXING FOR RESEARCH BASED INFORMATION RETRIEVAL*", has been conducted and concluded.

The finalised Volumes I and II were submitted to Kyle on 5 November 2019 and cc'd to Prof André de la Harpe.

**Sincerely**



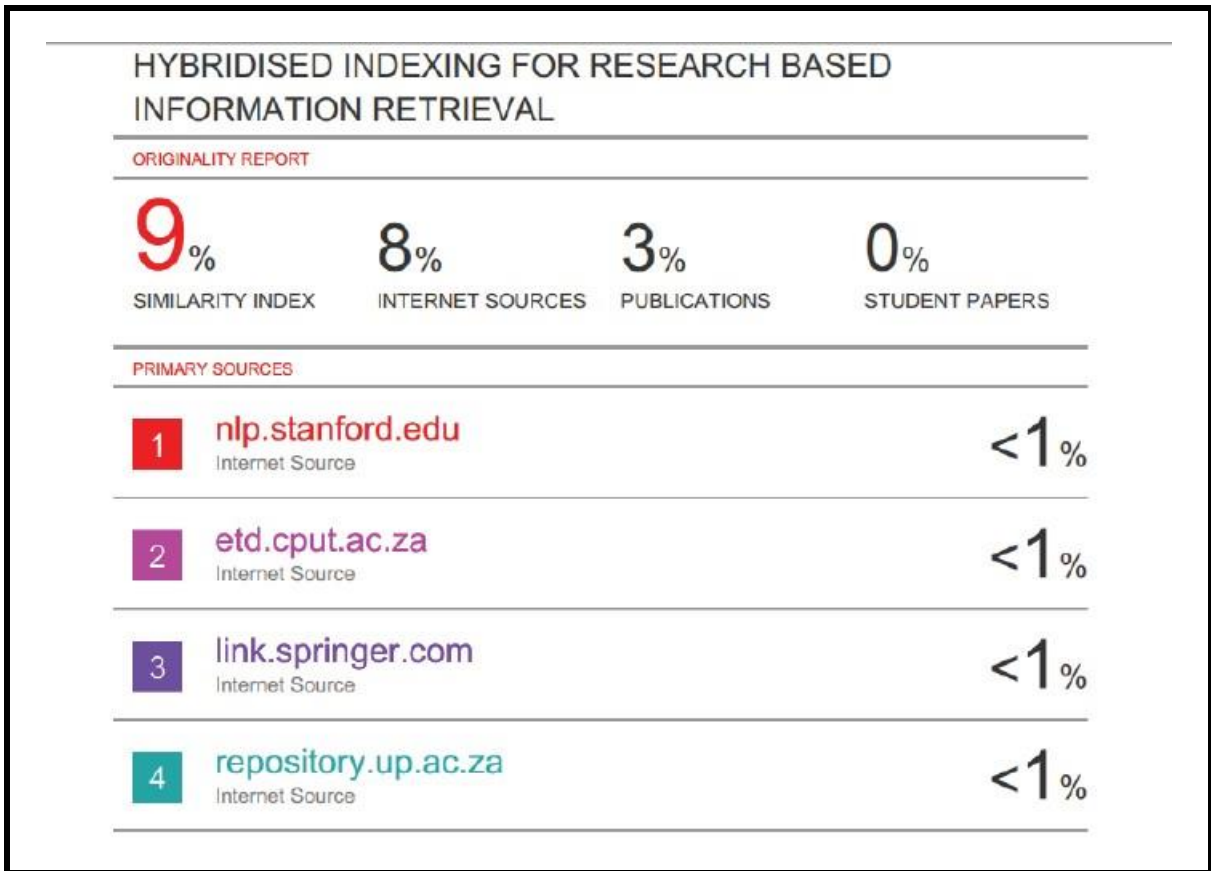
**Professor Annelie Jordaan**  
*D'Tech: Information Technology*  
*Ph: 061 420 9002*

**Member: SATI 1003347**



South African Translators' Institute (SATI)

## TURNITIN REPORT



## ABSTRACT

Challenges exist for information retrieval systems in handling mismatching vocabularies in queries and candidate source documents. As a result, these information retrieval systems may retrieve some documents that are non-relevant and miss some that are relevant. This increases the time for research by forcing additional perusal of unsatisfactory results, and additional searches using alternative vocabularies, which renders information retrieval systems less effective than they could be, and inhibits productive research.

The aim of this research was to design, build, and rigorously pilot test a hybrid indexing method that maintains phrase-term word ordinality and word proximity, and to compare the effectiveness of this method with the traditional inverted indexing method. The objectives were to prove statistically that the hybrid indexing method: i) increases the effectiveness of retrieving only those documents that are judged relevant by the user; ii) reduces errors in incorrect identification of user judged relevant documents, thus reducing the number of documents for the user to peruse; and iii) increases the rejection quality of user non-relevant documents, thus providing confidence to the user in the judgement of the information retrieval system. Finally, to determine whether this hybrid indexing method solves the problem of mismatching vocabulary between a query and a document, and satisfies the information needs of the user by retrieving only those documents from the collection relevant to the user. It must be noted that the results from the statistical analysis in this research are not the contribution to knowledge, as the statistics are used to prove that the hybrid indexing method worked. This indexing method is the contribution to the body of knowledge.

The strategy used was based on design science research performing both an exploratory and an explanatory study. Quantitative data were collected from the results of processing search queries through two information retrieval systems (one using the hybrid indexing method and the other the inverted indexing method) and from the results of a questionnaire completed by five participants during an experiment. The quantitative data were converted to binary and tested statistically using the mean averages for precision, recall, and specificity, and the Kappa coefficient.

The hybrid indexing method was presented and proved, with significance, to increase system effectiveness and specificity. Based on the results, the vocabulary mismatch problem between a query and a document was solved, but the information needs of the user were not satisfied.

**Keywords:** Hybrid token index, hybrid query index, research, information retrieval system, vocabulary mismatch, phrase-term frequency, unique token identity number, precision, recall, specificity

## **ACKNOWLEDGEMENTS**

I wish to sincerely thank:

- My wife Jane and my three children Claire, Johnathan, and Francesca for their understanding, patience, and endurance throughout this research
- Professor Andre de la Harpe for his supervision, dedication and support during this work
- Professor Andy Bytheway for his positivism and assistance during this incredible journey
- Doctor Corrie Uys for her professional statistical assistance and expertise
- Ann Bytheway, for proofreading this thesis
- Professor Retha de la Harpe, for all her support
- Professor Annelie Jordaan for editing this thesis
- Jay Barnes, Professor Justine Daramola and Doctor Boniface Kabaso for their guidance
- To the five participants who assisted in the experiment at CPUT

## **DEDICATION**

This thesis is dedicated to all children with Type 1 Diabetes Mellitus and to the parents who are searching for an answer to the root cause, together for a cure, for this dreadful disease.

## TABLE OF CONTENTS

<b>DECLARATION.....</b>	<b>ii</b>
<b>EDITOR’S CERTIFICATE .....</b>	<b>iii</b>
<b>TURNITIN REPORT .....</b>	<b>iv</b>
<b>ABSTRACT .....</b>	<b>v</b>
<b>ACKNOWLEDGEMENTS.....</b>	<b>vi</b>
<b>DEDICATION .....</b>	<b>vii</b>
<b>LIST OF FIGURES .....</b>	<b>xiv</b>
<b>LIST OF TABLES.....</b>	<b>xvi</b>
<b>LIST OF EQUATIONS.....</b>	<b>xvii</b>
<b>GLOSSARY .....</b>	<b>xix</b>
<b>CHAPTER ONE: INTRODUCTION AND BACKGROUND .....</b>	<b>1</b>
1.1 Introduction .....	1
1.2 Background to the research problem .....	5
1.3 The problem statement .....	7
1.4 Research questions and hypotheses .....	8
1.5 Aim of study .....	10
1.6 Research design .....	10
1.7 Data collection and analysis.....	11
1.7.1 Data collection .....	11
1.7.2 Data analysis .....	12
1.7.3 Statistical analysis.....	12
1.8 Findings .....	12
1.9 Recommendations for further research .....	13
1.10 Ethics.....	13
1.11 Delineation and limitations .....	14
1.12 Contributions.....	14
1.12.1 Theoretical contributions .....	14
1.12.2 Methodological contributions.....	14
1.12.3 Practical contributions .....	15
1.13 Summary .....	15
1.14 The framework of the thesis .....	17
<b>CHAPTER TWO: LITERATURE REVIEW.....</b>	<b>19</b>
2.1 Introduction .....	19



---

2.2	A brief history of information retrieval .....	21
2.3	Concepts of vocabulary mismatch.....	26
2.4	A user's pursuit for documents.....	31
2.5	IRS models and methods.....	36
2.5.1	The bag of words model.....	37
2.5.2	The citation ranking method.....	37
2.5.3	The Boolean retrieval model .....	37
2.5.4	The Vector space model .....	38
2.5.5	Markov random field model.....	38
2.6	IRS Indexing methods.....	39
2.6.1	The inverted index .....	41
2.6.2	The tiered index .....	43
2.6.3	The phrase index .....	44
2.6.4	The positional index .....	45
2.6.5	The next word index.....	45
2.7	IRS design concepts .....	47
2.7.1	Content acquisition .....	47
2.7.2	Text transformation .....	52
2.7.3	The data store.....	57
2.7.4	The token index .....	57
2.8	Search engines and queries .....	57
2.8.1	Query expansion.....	57
2.8.2	User relevance feedback .....	58
2.8.3	Ranked retrieval.....	59
2.8.4	Word proximity and ordinality .....	60
2.9	IRS matching and processing .....	63
2.9.1	The term-by-document matrix .....	64
2.9.2	Term frequency.....	64
2.9.3	Document frequency .....	65
2.9.4	Collection frequency .....	65
2.9.5	Documents in a collection .....	66
2.9.6	Inverse document frequency .....	66
2.9.7	Document weight .....	66
2.9.8	Cosine similarity theory .....	67
2.10	Measurements and formulae.....	69
2.10.1	Precision .....	72
2.10.2	Recall.....	73
2.10.3	F-measure .....	73

---

2.10.4	Accuracy .....	74
2.10.5	Snobbery ratio .....	75
2.10.6	Noise factor.....	75
2.10.7	Fallout.....	75
2.10.8	Specificity.....	76
2.10.9	Measurements of agreement .....	76
2.10.10	Scale of agreed judgments .....	78
2.10.11	IRS significance tests .....	78
2.11	The theoretical conceptual framework.....	79
2.11.1	The user stage.....	82
2.11.2	The IRS stage.....	82
2.11.3	The evaluation stage.....	83
2.12	Summary .....	83
<b>CHAPTER THREE: RESEARCH DESIGN.....</b>		<b>88</b>
3.1	Introduction .....	88
3.2	Research purpose.....	91
3.2.1	The exploratory study .....	92
3.2.2	The descriptive study .....	94
3.2.3	The explanatory study.....	94
3.3	Research philosophy .....	95
3.3.1	Ontology .....	96
3.3.2	Epistemology .....	98
3.3.3	Axiology .....	101
3.4	Research design.....	101
3.5	Research strategy.....	105
3.5.1	Design science research strategy .....	105
3.5.2	The experiment.....	109
3.5.3	Research approach.....	113
3.5.4	Research method.....	113
3.5.5	Research choice .....	113
3.5.6	Research time horizon .....	114
3.6	Data collection .....	114
3.6.1	Units of analysis and observation.....	114
3.6.2	Sampling techniques.....	115
3.6.3	The experiment.....	116
3.7	Data analysis .....	118

---

3.7.1	Performance measurements.....	118
3.7.2	Statistical analysis.....	118
3.8	Ethical considerations.....	119
3.9	Summary.....	119
<b>CHAPTER FOUR: RESEARCH RESULTS.....</b>		<b>121</b>
4.1	Introduction.....	121
SECTION A – THE EXPLORATORY STUDY.....		123
4.2	DSR – an introduction.....	123
4.3	The design of the hybrid index.....	123
4.4	The pilot tests.....	125
4.4.1	Pilot 1: Hamlet.....	125
4.4.2	Pilot 2: Ulysses.....	127
4.4.3	Pilot 3: Vocabulary mismatch.....	129
4.5	Designing and building the hybrid index.....	131
4.5.1	The information gathering process.....	131
4.5.2	The search engine process.....	133
4.6	Summary of hybrid indexing design findings.....	137
4.7	The first research question.....	138
SECTION B – THE EXPLANATORY STUDY.....		138
4.8	The experiment – an introduction.....	138
4.9	The experiment.....	139
4.9.1	The User.....	139
4.9.2	IRS-H.....	140
4.9.3	IRS-I.....	140
4.10	Hypothesis 1: Analysing effectiveness in retrieving relevant documents.....	140
4.10.1	Precision measurements.....	141
4.10.2	Ranking.....	141
4.10.3	Average precision measurements.....	143
4.10.4	Mean average precision.....	144
4.10.5	Student’s t-test and t-distribution.....	145
4.11	Hypothesis 2: Analysing incorrect identification of relevant documents.....	146
4.11.1	Recall measurements.....	146
4.11.2	Ranking.....	147
4.11.3	Average recall measurements.....	149
4.12	Hypothesis 3: Analysing rejection quality of non-relevant documents.....	151
4.12.1	Specificity measurements.....	152

---

4.12.2	Ranking .....	152
4.12.3	Average specificity measurements.....	154
4.12.4	Mean average specificity.....	155
4.12.5	Student's t-test and t-distribution.....	156
4.13	Hypothesis 4: Judgments made by IRS-H and the user .....	157
4.14	Hypothesis 5: Satisfying the user's information need .....	161
4.15	Summary of experimental findings .....	162
4.16	The second research question.....	163
4.17	Summary .....	163
<b>CHAPTER FIVE: DISCUSSION .....</b>		<b>165</b>
5.1	Introduction .....	165
5.2	The research design .....	166
5.2.1	The research problem.....	166
5.2.2	The aim of the research .....	168
5.2.3	The objectives of the research .....	168
5.2.4	The first research question.....	169
5.2.5	The hypotheses .....	169
5.2.6	The second research question.....	170
5.3	Findings: the proposed framework .....	171
5.3.1	The User stage .....	174
5.3.2	The IRS stage.....	174
5.3.3	The Evaluation stage .....	175
5.4	Findings: IRS-H versus IRS-I evaluation .....	175
5.4.1	Increasing effectiveness in retrieving relevant documents.....	176
5.4.2	Reducing incorrect identification of relevant documents.....	176
5.4.3	Increasing quality in rejecting non-relevant documents.....	176
5.5	Findings: IRS-H versus User judgements.....	177
5.6	Findings: IRS-H versus User results .....	177
5.7	The second research question .....	178
5.8	Summary of the framework.....	181
<b>CHAPTER SIX: CONCLUSIONS AND RECOMMENDATIONS .....</b>		<b>183</b>
6.1	Introduction .....	183
6.2	Conclusions .....	183
6.2.1	The first research question.....	183

---

6.2.2	Hypotheses tested: IRS-H versus IRS-I .....	184
6.2.3	Hypotheses tested: IRS-H versus User .....	185
6.2.4	The second research question .....	185
6.2.5	Aim of the study .....	186
6.2.6	General .....	188
6.2.7	Summary .....	188
6.3	Recommendations for this research .....	189
6.3.1	From a user's perspective .....	189
6.3.2	From an IRS perspective .....	189
6.3.3	From an evaluation perspective .....	190
6.4	Recommendations for further research .....	190
6.5	Summary .....	192
 <b>CHAPTER SEVEN: CONTRIBUTIONS AND REFLECTIONS .....</b>		<b>193</b>
7.1	Introduction .....	193
7.2	Contributions .....	193
7.2.1	Theoretical contributions .....	193
7.2.2	Methodological contributions .....	194
7.2.3	Practical contributions .....	194
7.3	Reflection .....	197
7.3.1	Assessment of design .....	197
7.3.2	Assessment of contributions .....	199
7.3.3	Assessment of the research .....	199
7.3.4	Assessment of the context and research purpose .....	200
7.3.5	Self-reflection .....	200
 <b>REFERENCE LIST .....</b>		<b>202</b>

**APPENDICES are submitted separately in Volume II**

## LIST OF FIGURES

Figure 1.1: Schematic representation of Chapter One .....	1
Figure 1.2: A simple flow diagram of the research design .....	11
Figure 1.3: The framework of this thesis .....	17
Figure 2.1: Schematic representation of Chapter Two .....	19
Figure 2.2: Markov Random Field model assumptions.....	39
Figure 2.3: Building an index.....	42
Figure 2.4: A sample next word index.....	46
Figure 2.5: Estimated size of Google and Bing indices .....	49
Figure 2.6: A sample of papers from the NIST 2018 collection.....	50
Figure 2.7: An ACM collection.....	51
Figure 2.8: Representation of sentences.....	54
Figure 2.9: A document text example.....	63
Figure 2.10: A term-by-document matrix.....	65
Figure 2.11: A simple feedback scheme .....	79
Figure 2.12: A theoretical conceptual framework from the literature.....	82
Figure 3.1: Schematic representation of Chapter Three.....	88
Figure 3.2: Theoretical conceptual framework with research question and hypotheses .....	91
Figure 3.3: A flow chart representing the exploratory study.....	92
Figure 3.4: A flow chart representing the explanatory study.....	94
Figure 3.5: The research onion .....	96
Figure 3.6: A three-world ontology .....	97
Figure 3.7: Four paradigms for the analysis of social theory .....	98
Figure 3.8: A simple flow diagram for this research.....	103
Figure 3.9: A graphic illustration of the research design for this study.....	104
Figure 3.10: Design science research cycles .....	106
Figure 3.11: The three design cycles for this study .....	108
Figure 3.12: An example of the questionnaire .....	111
Figure 3.13: A flow chart representing system-generated data .....	112
Figure 3.14: A flow chart representing system and user-generated judgments .....	113
Figure 3.15: Sampling techniques adopted for this research.....	116
Figure 3.16: The experimental framework.....	117
Figure 4.1: Schematic representation of Chapter Four.....	121
Figure 4.2: The three artefacts of IRS-H .....	124
Figure 4.3: The two artefacts of IRS-I .....	124
Figure 4.4: The design of the hybrid index .....	125
Figure 4.5: Pilot 1 design changes.....	127

---

Figure 4.6: Pilot 2 design changes .....	129
Figure 4.7: Pilot 3 design changes .....	130
Figure 4.8: The information gathering process .....	131
Figure 4.9: An example of content acquisition and text transformation.....	132
Figure 4.10: Entity relationship diagram for hybrid token indexing.....	133
Figure 4.11: The search engine process .....	133
Figure 4.12: Entity relationship diagram for hybrid query indexing .....	136
Figure 4.13: The experimental framework.....	139
Figure 4.14: IRS-I query 1 – Precision values for one ranking of 23 relevant documents ...	142
Figure 4.15: IRS-H query 1 – Precision values for one ranking of 13 relevant documents .	143
Figure 4.16: Mean average precision t-distribution and results .....	145
Figure 4.17: IRS-I query 1 – Recall values for one ranking of 23 relevant documents.....	148
Figure 4.18: IRS-H query 1 – Recall values for one ranking of 13 relevant documents .....	148
Figure 4.19: Mean average recall t-distribution and results .....	151
Figure 4.20: IRS-I query 1 – Specificity values of 26 rejected non-relevant documents.....	153
Figure 4.21: IRS-H query 1 – Specificity values of 66 rejected non-relevant documents ....	154
Figure 4.22: Mean average specificity t-distribution and results .....	156
Figure 5.1: Schematic representation of Chapter Five .....	165
Figure 5.2: The layout of Chapter Five .....	166
Figure 5.3: The research problem and aims.....	167
Figure 5.4: The proposed framework of the hybrid indexing method .....	172
Figure 6.1: Schematic representation of Chapter Six .....	183
Figure 6.2: The proposed framework of the hybrid indexing method .....	187
Figure 7.1: Schematic representation of Chapter Seven .....	193

## LIST OF TABLES

Table 1.1: The hypotheses, groups, and variables.....	8
Table 1.2: Research questions, hypotheses, objectives and methods .....	9
Table 2.1: An early document-by-citation year matrix.....	33
Table 2.2: Journal citation frequencies.....	34
Table 2.3: Data retrieval vs. information retrieval .....	35
Table 2.4: A confusion matrix.....	70
Table 2.5: A 2x2 contingency table (a,b,c,d) .....	71
Table 2.6: A 2x2 contingency table .....	71
Table 2.7: The stages and key concepts derived from the literature.....	80
Table 3.1: The design choices .....	104
Table 3.2: Control and test groups: H1, H2 and H3.....	111
Table 3.3: Control and test groups: H4 and H5 .....	112
Table 4.1: Research questions, hypotheses, objectives, methods and sections.....	121
Table 4.2: An example of the hybrid token index.....	132
Table 4.3: An example of the hybrid query index .....	134
Table 4.4: Summary of hybrid indexing design findings .....	137
Table 4.5: Results of average precision measurements per query per indexing method ....	143
Table 4.6: Average recall measurements per query per indexing method .....	149
Table 4.7: Average specificity measurements per query per indexing method .....	154
Table 4.8: Kappa coefficient agreement measures .....	158
Table 4.9: IRS-H versus user judgements – phrase-terms.....	159
Table 4.10: IRS-H versus User – mismatched judgement cases.....	160
Table 4.11: User versus IRS-H – mismatched judgement cases.....	160
Table 4.12: IRS-H versus User judgements – information needs .....	161
Table 4.13: Summary of experimental findings .....	162
Table 5.1: Design findings of the hybrid indexing method .....	173
Table 5.2: Summary of findings .....	182



## LIST OF EQUATIONS

Equation 2.1: Weight.....	66
Equation 2.2: Cosine similarity $\text{sim}(d3,d4)$ .....	68
Equation 2.3: Cosine similarity $\text{sim}(q1,d4)$ .....	68
Equation 2.4: Cosine similarity $\delta_i(q1,d4)$ .....	68
Equation 2.5: Cosine similarity $\delta_4(q1,d4)$ .....	69
Equation 2.6: Precision .....	72
Equation 2.7: Precision .....	72
Equation 2.8: Precision .....	73
Equation 2.9: Recall.....	73
Equation 2.10: Recall.....	73
Equation 2.11: Recall.....	73
Equation 2.12: F .....	74
Equation 2.13: F1 .....	74
Equation 2.14: F-measure .....	74
Equation 2.15: Accuracy (a,b,c,d) .....	75
Equation 2.16: Accuracy (tp,fp,fn,tn) .....	75
Equation 2.17: Snobbery ratio.....	75
Equation 2.18: Noise factor.....	75
Equation 2.19: Fallout .....	76
Equation 2.20: Specificity.....	76
Equation 2.21: P(A) .....	77
Equation 2.22: P(non-relevant) .....	77
Equation 2.23: P(relevant) .....	77
Equation 2.24: P(E) .....	77
Equation 2.25: k.....	78
Equation 4.1: P .....	141
Equation 4.2: PIRS-Iq01 .....	141
Equation 4.3: PIRS-Hq01.....	141
Equation 4.4: MAP .....	144
Equation 4.5: MAPIRS-I.....	144
Equation 4.6: MAPIRS-H .....	144
Equation 4.7: R.....	146
Equation 4.8: RIRS-Iq01 .....	147
Equation 4.9: RIRS-Hq01 .....	147
Equation 4.10: MAR (Zhao & Huang, 2016:3).....	150
Equation 4.11: MARIRS-I.....	150

Equation 4.12: MARIRS-I.....	150
Equation 4.13: S .....	152
Equation 4.14: SIRS-Iq01 .....	152
Equation 4.15: SIRS-Hq01.....	152
Equation 4.16: MAS .....	155
Equation 4.17: MASIRS-I.....	155
Equation 4.18: MASIRS-H .....	155

## GLOSSARY

Term	Explanation
A	Accuracy
AI	Artificial Intelligence
AP	Average precision
AR	Average recall
AS	Average specificity
BoW	Bag of words
cf	Collection frequency – the number of times a token occurs within a document collection
Corpus	A large set of texts usually electronically stored and processed
Corpora	More than one corpus
CPUT	Cape Peninsula University of Technology
CSV	Comma Separated Value (a file format)
d	Document
df	In information retrieval: document frequency – the number of documents in which a term or phrase-term occurs
df	In statistics: the degrees of freedom
DRC	Democratic Republic of the Congo
DSR	Design Science Research
Effective	Successful in producing the desired result
ERD	Entity Relationship Diagram
F	F-measure – a measure of the overall effectiveness of an information retrieval system
fn	False negative – the number of user relevant documents not retrieved by the system
fntn	False negative and true negative – the number of documents not retrieved by the system
fo	Fallout ratio
fp	False positive – the number of user non-relevant documents retrieved by the system
fptn	False positive and true negative – the number of user non-relevant documents
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ICT	Information and Communications Technology

Term	Explanation
idf	Inverse document frequency
IR	Information retrieval – is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)
IRS	Information Retrieval System
IRS-H	An information retrieval system using the hybrid indexing method
IRS-I	An information retrieval system using the inverted indexing method
k	Kappa coefficient
MAP	Mean average precision
MAR	Mean average recall
MAS	Mean average specificity
MS-Access	Microsoft Access – a database
N	Document collection – the number of documents in the collection (the sum of tp + fp + fn + tn)
Nf	Noise factor – the measure of degradation within a system
Non-relevant	In information retrieval the term 'non-relevant' is used to describe the antonym of 'relevant' rather than the more appropriate English language terms of 'irrelevant' and 'not relevant'
OCR	Optical Character Recognition
P	Precision – the ratio of the number of relevant documents retrieved and the number of documents retrieved
P	In statistics: the p significance level
pdf	Portable Document Format (a file format)
Phrase query	A query containing one or more phrase-terms
pt	Phrase-term – a multi-word term used to describe a concept
ptf	Phrase-term frequency – the number of times a phrase-term occurs within a document. It is a key measurement in evaluating the hybrid indexing method
q	Query – a question requesting information
R	Recall – the ratio of the number of relevant documents retrieved and the number of relevant documents in a collection
Relevance	The concept of similarity where a query is connected to a document
S	Specificity – the quality in rejecting non-relevant documents
S4HANA	SAP for Hana – SAP AGs business application that uses in-memory database computing and cloud based computing
SA	South Africa

Term	Explanation
Sn	Snobbery ratio – the complement to Recall – the ratio of the number of relevant documents not-retrieved and the number of relevant documents in a collection
SPSS	IBM SPSS statistics version 25
SQL	Structured Query Language
t	In information retrieval: a term – a word used to describe a concept
t	In statistics: the t-value
tf	Term frequency – the number of times a term occurs within a document
tf*idf	The product of term frequency (tf) and inverse document frequency (idf) – alternatively represented as tf_idf
tn	True negative – the number of user non-relevant documents not retrieved by the system
Token	A chunk of data acquired from text
Token ID	Unique token identity number – a key design concept for both the hybrid token index and the hybrid query index
tp	True positive – the number of user relevant documents retrieved by the system
tpfn	True positive and false negative – the number of user relevant documents
tpfp	True positive and false positive – the number of documents retrieved by the system
txt	Text (a file format)
UCT	University of Cape Town
UoA	Unit of analysis – in this research it is a query
UoO	Unit of observation – in this research it is a document
USA	Unites States of America
US-English	Unites States English language – spelling differs from the British English language
User	User refers to participants answering a questionnaire during an experiment, and user refers to a system when data are acquired from the completed questionnaire
VB	Visual Basic – a programming language
Vocabulary mismatch	The mismatch that occurs between terms expressed in a query and words within a document
Windows	Windows version 10 – an operating system
Word ordinality	A number indicating the position of a word in a sentence

Term	Explanation
Word proximity	Two or more words are within a specified distance – distance is the number of intermediate words
Word term	A single-word term
$\alpha$	In statistics: the $\alpha$ significance level
$\omega$	Word proximity

## CHAPTER ONE: INTRODUCTION AND BACKGROUND

*“In research, the present devours the past” – Medawar*

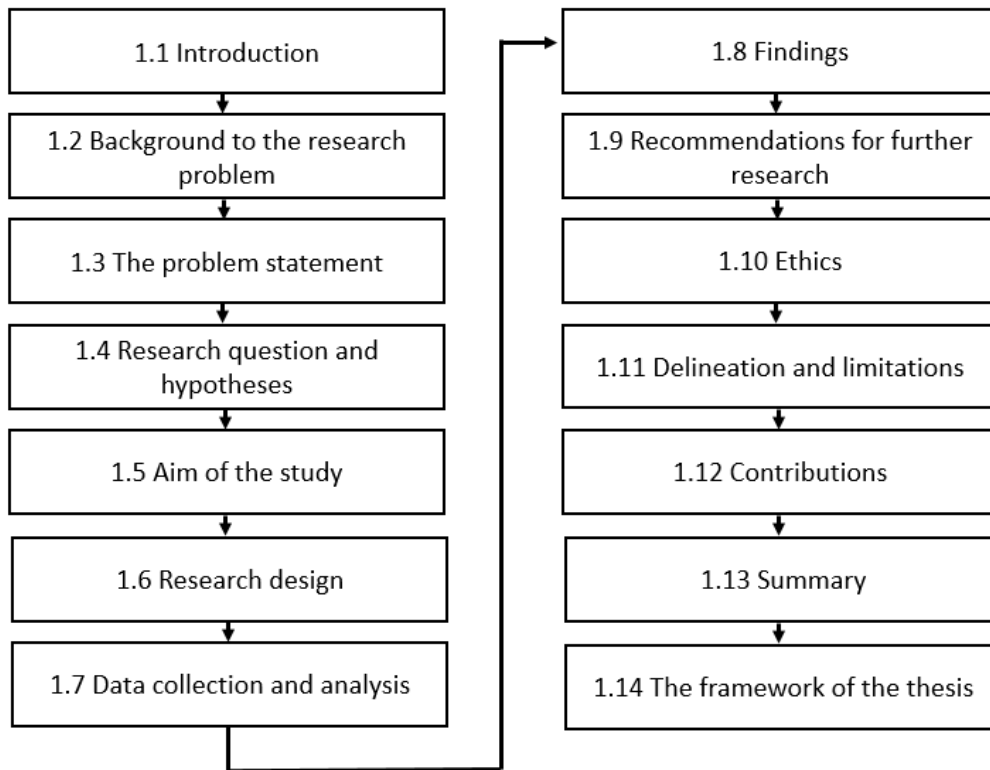


Figure 1.1: Schematic representation of Chapter One

### 1.1 Introduction

During the early stages of this researcher’s search for the literature for his master’s thesis, it became apparent on many occasions that the sets of documents searched for were non-relevant to the required information needs. These non-relevant documents prompted an enquiry into developing a method to retrieve information during the research process better.

Searching for information should be easier and more efficient than in the recent past. It has become common for work colleagues, friends, and family members to say, “You do not know? Then just Google it! All the information is there at your fingertips”. But this is not so.

During the early stages of this research, trying to find documents that pertained to what was being searched for became frustrating even with the World Wide Web (Web) search engines available today. Many search engines were used during that master’s research, including Google Search, Google Scholar, Yahoo, Anansi, and Bing among

others, but frequently they returned disappointing results in relation to the information needed. On a few occasions the results returned were of merit, at other times they were average and many times far too many documents were returned that were totally irrelevant. This sort of situation creates a tedious task for the researcher who has to read through and then reject the irrelevant documents. On many occasions, documents returned are copyrighted by the journal hosting the article on the Web and a fee is requested to allow a download to one computer. However, this becomes financially risky, as it is impossible to confirm, at this stage, that the article contains relevant and needed information. Consequently, university libraries were turned to in the hope of seeking and retrieving the relevant documents. Online libraries at the Cape Peninsula University of Technology (CPUT) and the University of Cape Town (UCT) were explored using many title and keyword searchers. This became equally frustrating, as many of the articles available on the Web (for a fee) were unavailable in these libraries; and for those that were in the libraries, searching for them became challenging as the user had to use the correct keywords, although unknown at that point in time, that were registered in the various databases associated with the document sought. This problem of mismatching was compounded when multi-word phrase-terms were used increasing the chance of irrelevant documents being returned, because many of these information retrieval systems (IRSs) could not maintain word ordinality<sup>1</sup> and word proximity<sup>2</sup> for the words used in the phrase-terms. If a two-word phrase-term was used, there was no guarantee that the two words existed side by side in the document returned, as one word could appear on the first page and the other on the last page. However, it is noted that operational IRSs have provided word proximity methods since the 1980s, for example, in services and databases such as Dialog (Anon, 2006), DataStaR (Khan, 2010) and EBSCO (EBSCO, 2019).

There were 'Find' options available in various software applications that read Word format documents and Portable Document Format (PDF) documents. Adobe's PDF "Find" was more sophisticated than that of Microsoft Word but still did not produce the desired results. The inadequacies of the existing software applications (Koopman & Zuccon, 2019) triggered this researcher to design a method in an attempt to solve the problem of mismatching terms through poor word ordinality and word proximity. To execute this method, a prototype IRS had to be built that made searching possible by combining many multi-word terms expressed in many queries attempting to satisfy

---

<sup>1</sup> A number indicating the position of a word in a sentence

<sup>2</sup> Two or more words are within a specified distance (distance is the number of intermediate words)



the researcher's information needs. These searches were expanded to retrieve those research documents that pertained to a specific research methodology (e.g. design science research), and a specific ontological stance (e.g. post-positivism). One specific need was to find recent articles, from specific journals, that described the problem of mismatching terms used in a search query and the terms used in that document. Theory surrounding this problem is referred to as vocabulary mismatch, and this vocabulary mismatch is a synonymic example of itself as it has been described in many differing ways, for example, vocabulary mismatch, term mismatch, vocabulary problem, and vocabulary gap. Therefore, in the prototype, specific queries were designed to use many multi-word phrase-terms to describe the same concept. For the system to accommodate this requirement, changes to the IRS indexing design became necessary.

IRSs together with their search engines have become synonymous with the Web in retrieving the documents sought by a user. These challenges however do not only apply to the Web, but also to other smaller closed collections that exist for libraries, journals, and many researchers' own academic literature. When using an IRS, words of a language are used to express an information need in the form of a query to find what they want and thereafter they perform a search in an attempt to retrieve that information.

With that said, it is important at this early stage to position this research for the reader. Currently there are Web search engines and Web IRSs that assist researchers and others to retrieve information from the Web. Examples of search engines are Google, Yahoo, Anansi, Bing, and Baidu. When using Web search engines, the user is unaware of which document collection the search engine is using and which Web pages are referenced within the database. For all intended purposes, the databases that Web search engines use can be seen as a black box because the content of the database is unknown to the user. Search engines rely on the effectiveness of information gathering by Web crawlers and/or spiders crawling the Web to gather information from Web pages and textual documents. It is this gathering of information that is used to populate the database readying the information for searching within the IRS. A search engine within an IRS can only use the information that the IRS has gathered and hence the reasons for a user's preferential use (for example, the database content is more country specific or the search query is more precise) of the various Web search engines available today.

This research is not about designing and building another Web search engine to assist those wanting to retrieve information from the Web. This research is positioned

around smaller non-Web-based document collections, for example, university libraries that contain many books in electronic format; various documents used by businesses (purchase orders, invoices); a researcher's own personal closed collection of documents acquired from journal articles, conference papers, theses; and other sources and documents acquired from professional online keyword based document databases including EBSCOhost, ProQuest, Emerald and Scopus.

In the business world, closed document collections are frequently used. An example, which was a stimulus for this research, was the divestiture by a mining company back in 2017. This example is now presented:

- i) Freeport-McMoRan, a United States of America (USA) company, sold its copper and cobalt mine Tenke Fungurume in the Democratic Republic of the Congo (DRC) to China Molybdenum Company Limited (CMOC). From an Information and Communications Technology (ICT) perspective, the project objective was to identify and then separate the mine's data and text information from the SAP for Hana (S4HANA) computer servers in the USA, and thereafter to migrate the data and text held in databases and closed document collections to cloud-based computer servers hosted in Johannesburg, South Africa (SA).
- ii) During migration of the over 10,000 documents, a data quality initiative was necessary to cleanse historical textual data held within the system. This initiative encompassed the use of specific vocabulary and the search for key phrases where word ordinality and word proximity were critical within the unstructured text of documents (purchase order details, invoice details, and material descriptions). The objective was to understand the content of these documents better, thus enabling the creation of metadata that could better describe the contents of these documents, and could provide a more effective search to recall the documents needed.
- iii) The challenge was to identify indexing methods and the phrases, keywords and the arrangement of words (using ordinality and proximity) that existed within the documents and to identify which of the documents contained the specific phrases.

The word 'indexing' refers to the way in which data stored in a computer can be retrieved in response to a query. The word 'indexing' also refers to the way in which the content of a document is represented. This representation can be performed using: terms from a controlled vocabulary, for example, thesaurus or assigned keywords or a classification scheme.

In this research, the number of documents in the researcher's closed collection is deemed smaller, hundreds or thousands of documents rather than the tens of thousands of company documents. Once a researcher has collected documents from whatever sources are of interest to him/her and these documents now reside on the researcher's computer, it is at this point that research begins.

To avoid confusion, the definition for information retrieval in this research is now presented. Manning, Raghavan and Schütze (2008:1) state that "information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)".

Based on this definition by Manning et al. (2008), this research aimed to design, build, and rigorously pilot test a hybrid indexing method that maintains phrase-term word ordinality and word proximity (IRS-H), and to compare the effectiveness of this method with the traditional inverted indexing method (IRS-I). In addition, the second aim was to design a method and build an IRS that solves the problem of mismatching vocabulary between a query and a document, and satisfies the information needs of the user.

## **1.2 Background to the research problem**

Vocabulary mismatch is a phenomenon whereby multiple words in a phrase used to describe something in the past change over time to describe the same thing (Shekarpour et al., 2017), and when these phrases are expressed within a search query a mismatch occurs between the query and the document (Onal et al., 2018). Vocabulary mismatch thus affects the effectiveness of text-based IRSs, as the words within the phrases, expressed as queries, are not accurately matched to the words within the text of documents within a collection (Nguyen et al., 2018).

According to Onal et al. (2018), the vocabulary of the user and the vocabulary in relevant documents may differ and therefore fewer relevant documents may be retrieved. This affects the searching for information in the legal world according to Andersson, Rekabsaz and Hanbury (2017), where patent text is generally a mixture of legal and domain specific terms often making use of synonyms. Goeuriot et al. (2016), Dietz et al. (2019), and Koopman and Zuccon (2019) concur that vocabulary mismatch affects healthcare searches, as often relevant documents are not retrieved by standard approaches. In commerce, Liu et al. (2017) argue the need for improvement to close the effect of vocabulary gap between product reviews and questions pertaining to these products. Shekarpour et al. (2017) argue that

vocabulary mismatch affects many professions as there is often a lack of accurate knowledge of the vocabulary used and even the experts frequently use specific vocabulary incorrectly.

After many attempts by various authors, the phenomenon of vocabulary mismatch in IRSs remains unresolved. For example, van Gysel, Li and Kanoulas (2018) apply two methods: a Bayesian optimisation intensified lexical method and a latent vector space method referred to as a neural vector space model, the latter having some degree of success. Referring to the vocabulary mismatch problem as the semantic gap, “car is a synonym of motorcar”, and to the relevancy of a user’s documents, Nguyen et al. (2018) explain that the semantic gap does hinder the matching of a query and a document and that this is the important problem to solve in order to select candidate relevant documents from a user's query.

In healthcare, Jimmy, Zucco and Koopman (2018) confirm that vocabulary mismatch remains a problem and argue that it is caused by the mismatch in the use of terms by the user and those used by the IRS. In their work, Jimmy et al. (2018) utilise a query feature expansion model using query expansion in an attempt to overcome the problem of poorly selected words expressed within a query. Koopman, Russell and Zucco (2018) concur with Pal, Mitra and Bhattacharya (2015) that vocabulary mismatch still remains a problem and state that this remains one of the important challenges when using keyword-based IRSs.

Vocabulary mismatch exists (Nguyen et al., 2018), as it affects the criteria of the search query for the user’s information need, where the IRS attempts to match the phrases within the query to those contained within a document via an index. These attempts remain a guess, as there is an element of uncertainty owing to the incompleteness of the query criteria (Van Rijsbergen, 1979). In his work at the time, Van Rijsbergen (1979) explains that in traditional data retrieval the query specification is complete as the information need is relatively precise, whereas in text retrieval the query specification is very often incomplete through uncertainty. Van Rijsbergen (1979) explains further that data retrieval efficiency is extremely strong as it generates an exact match making these systems highly efficient, whilst in text retrieval efficiencies are weak owing to this uncertainty.

Mitra and Awekar (2017) argue that vocabulary mismatch occurs through inexact term matches between the query and the document and thus the ineffectiveness in retrieving documents remains a problem. Mitra and Awekar (2017) and Shekarpour et al. (2017) hint at the requirement for exact phrase matching and a review of

indexing design. Although indexing methods in information retrieval using text are dealt with in detail in Chapter Two, a few concepts and applications of hybridised indexing are now discussed:

- i) Faloutsos and Jagadish (1992) describe a performance based hybrid approach to text indexing using dynamic databases and provide formulae and procedures on how to choose design parameters.
- ii) Navarro and Baeza-Yates (2001) use the phrase *hybrid indexing method* to describe a method for string matching where patterns are partitioned into chunks, these chunks are then searched for in a suffix tree, and finally, the positions of each chunk are verified to determine a complete match.
- iii) Ding, Li and Peng (2006) describe a hybrid indexing method that uses non-negative matrix factorisation and probabilistic latent semantic indexing successfully applied to document clustering.
- iv) Matveeva and Levow (2007) introduce a hybrid document indexing method that evaluates the prediction of topic boundaries based on chunks of text. This method uses spectral embedding that estimates semantic association between nouns over a distance of multiple chunks of text.
- v) Huang and Huang (2016) refer to a specific design of a hybrid index for non-parametric multivariate standardised drought indexing that considers variations in climatic precipitation and streamflow.

In summary, the vocabulary mismatch phenomenon remains a problem owing to the inability of, and IRSs indices forming the core workings of an IRS to interrogate each other effectively. From a search of the literature, it appears that a practical method that solves the vocabulary mismatch problem and reduces the retrieval of user unwanted non-relevant documents through the use of a novel indexing design, is still required. This research provides such an approach using a hybridised indexing method.

### **1.3 The problem statement**

Challenges exist for information retrieval systems in handling mismatching vocabularies in queries and candidate source documents (Onal et al., 2018). As a result, these information retrieval systems may retrieve some documents that are non-relevant and miss some that are relevant (Van Gysel, 2017). This increases the time of research by forcing additional perusal of unsatisfactory results, and additional searches using alternative vocabularies (Liu et al., 2017). This renders information retrieval systems less effective than they could be, and inhibits productive research (Mitra & Awekar, 2017; Nguyen et al., 2018).

#### 1.4 Research questions and hypotheses

The two research questions and the five hypotheses are presented in order to find solutions for the research problem by means of triangulation (Yeasmin & Rahman, 2012) and by defining the variables. The first research question is:

**RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

Table 1.1: The hypotheses, groups, and variables

Hypothesis	Control group	Test group	Independent variable	Dependent variable
H1	IRS-I	IRS-H	Hybridised indexing	Retrieval effectiveness
H2	IRS-I	IRS-H	Hybridised indexing	Incorrect identification of relevant documents
H3	IRS-I	IRS-H	Hybridised indexing	Quality in rejecting non-relevant documents
H4	User	IRS-H	The hybrid indexing method	Agreement in judgements
H5	User	IRS-H	The hybrid indexing method	Satisfying the information needs of the user

Table 1.1 presents the five hypotheses together with the control and test groups and the independent and dependent variables. For the first three hypotheses, the control group is IRS-I (using the inverted indexing method) and the test group is IRS-H (using the hybrid indexing method). The independent variable is hybridised indexing and the three dependent variables are: i) retrieval effectiveness; ii) incorrect identification of relevant documents; and iii) quality in rejecting non-relevant documents. For the final two hypotheses, the control group is the user (a group of participants) and the test group IRS-H. The independent variable is the hybrid indexing method and the two dependent variables are: i) agreement in judgements; and ii) satisfying the information needs of the user.

The second research question is:

**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

The research questions, hypotheses, objectives, and methods are summarised in Table 1.2.

Table 1.2: Research questions, hypotheses, objectives and methods

Research Question / Hypothesis	Aim / Objective	Method
<b>RQ1:</b> How can an IRS index be designed that maintains word ordinality and word proximity?	To design, build, and rigorously pilot test a hybrid indexing method that maintains word ordinality and word proximity, and to compare the effectiveness of this method with the traditional inverted indexing method	Literature review Exploratory Design science research Hybrid index design and build (IRS-H) Perform three pilot tests
<b>H1<sub>0</sub>:</b> Hybridised indexing does not increase the effectiveness of retrieving relevant documents	To test whether an IRS using a hybrid indexing method increases the effectiveness of retrieving only those documents that are judged relevant by the user	Literature review Explanatory, Experiment IRS-I and IRS-H tests Performance measurements Precision, Ranking, MAP Statistical analysis One-tailed t-test
<b>H2<sub>0</sub>:</b> Hybridised indexing does not reduce the incorrect identification of relevant documents	To test whether the hybrid indexing method reduces errors in incorrect identification of user judged relevant documents thus reducing the number of documents for the user to peruse	Literature review Explanatory, Experiment IRS-I and IRS-H tests Performance measurements Recall, Ranking, MAR Statistical analysis One-tailed t-test
<b>H3<sub>0</sub>:</b> Hybridised indexing does not increase the quality in rejecting non-relevant documents	To test whether the hybrid indexing method increases the rejection quality of user non-relevant documents thus providing confidence to the user in the judgement of the IRS	Literature review Explanatory, Experiment IRS-I and IRS-H tests Performance measurements Specificity, Ranking, MAS Statistical analysis One-tailed t-test
<b>H4<sub>0</sub>:</b> Judgments made by the hybrid indexing method and the user disagree	To determine whether the judgments made by the hybrid indexing method and the user agree	Literature review Explanatory, Experiment User judgements IRS-H judgements Kappa coefficient Agreement measurements
<b>H5<sub>0</sub>:</b> The hybrid indexing method does not satisfy the information needs of the user	To determine whether the hybrid indexing method satisfies the information needs of the user by retrieving those documents from the collection that are relevant to the user	Literature review Explanatory, experiment User judgements IRS-H judgements Kappa coefficient Agreement measurements
<b>RQ2:</b> Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?	To determine whether the hybrid indexing method solves the problem of mismatching vocabulary between a query and a document	Literature review Exploratory and Explanatory results from RQ1 and H1, H2, H3, H4 and H5 and findings

### **1.5 Aim of study**

The first aim of this research was to perform an exploratory study using design science research (DSR), by designing, building and then rigorously pilot testing a new IRS using a hybrid indexing method that maintains word ordinality and word proximity. The second aim was to perform an explanatory study, via experimentation, by comparing the effectiveness of the hybrid indexing method with that of the traditional inverted indexing method. Thereafter, to prove statistically that the hybrid indexing method increases the effectiveness of retrieving only those documents that are judged relevant by the user; reduces errors in incorrect identification of user judged relevant documents, thus reducing the number of documents for the user to peruse; and increases the rejection quality of user non-relevant documents, thus providing confidence to the user in the judgement of the IRS. Finally, to determine whether this hybrid indexing method solves the problem of mismatching vocabulary between a query and a document, and satisfies the information needs of the user by retrieving only those documents from the collection relevant to the user. The results from the statistical analysis in this research are used to prove that the hybrid indexing method works. This indexing method is the contribution to knowledge in this research.

### **1.6 Research design**

The design for this research encompassed stating the research problem, followed by a comprehensive literature review. The design for this research was dual purpose, as illustrated in Figure 1.2, by performing:

- i) An exploratory study based on DSR, and
  - to design and physically build an IRS,
  - to design and build a new indexing method utilising a pair of hybrid indices,
  - to review the literature numerous times gaining insight into existing theories, and
  - to perform pilots tests using various text based document collections.
- ii) An explanatory study by conducting an experiment:
  - where phrase-terms are expressed as queries,
  - the phrase-terms are applied to the IRS search engine,
  - the phrase-terms are applied to the user questionnaire,
  - both indexing methods are tested, the hybrid (IRS-H) and inverted (IRS-I), and
  - comparisons are made between the data generated from these indexing methods.



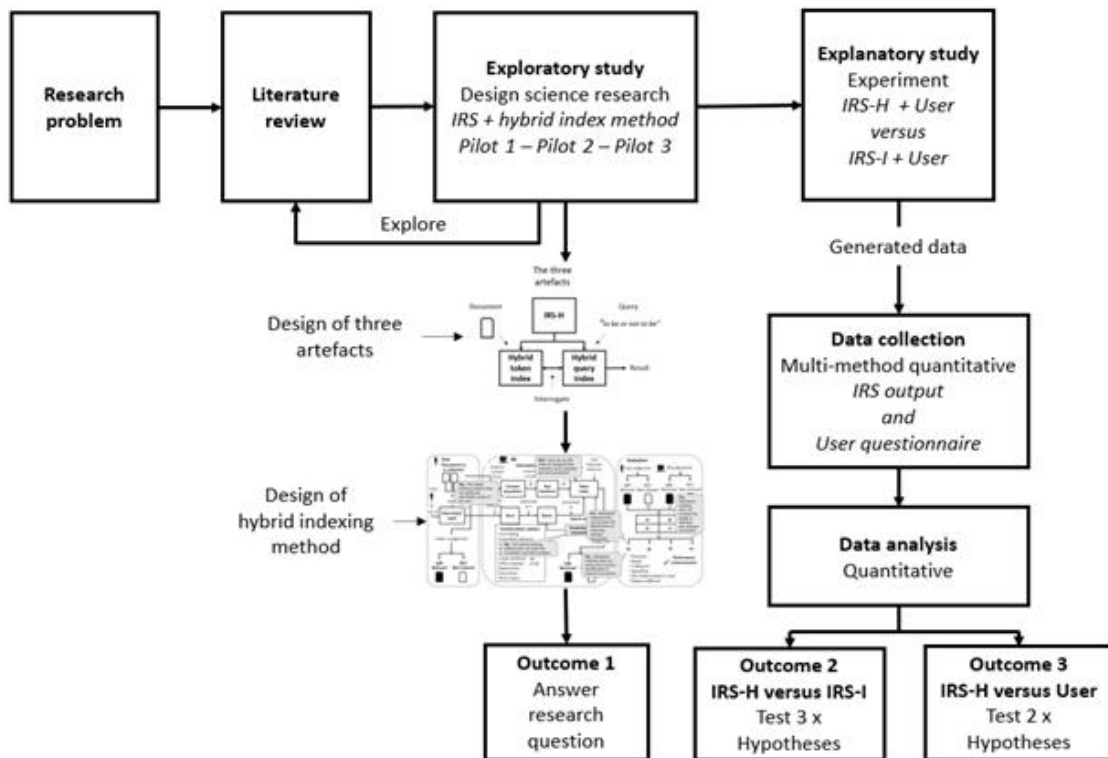


Figure 1.2: A simple flow diagram of the research design

Data collection (section 3.6) was thus multi-method quantitative to accommodate both research strategies of, an exploratory study using DSR (IRS), and an explanatory study using experimentation (IRS and questionnaire). The outcomes from DSR were to answer the first research question relating to IRS design, and the outcomes from experimentation using the quantitative data analysis (section 3.7) were threefold: i) to compare IRS-H with IRS-I and test three hypotheses; ii) to compare IRS-H with the user and test two hypotheses; and iii) to answer the second research question.

## 1.7 Data collection and analysis

### 1.7.1 Data collection

Saunders, Lewis and Thornhill (2016, 2019) suggest that with positivism, data collection should be highly structured using large samples together with specific measurements and quantitative data. In this research, the data collection approach was highly structured, as most of the data were system generated by two IRSs. However, data collection for the user's judgements was performed manually from five participating users after completing a predefined questionnaire pertaining to a set of queries. System-generated quantitative data were collected from the search engine results produced by the two IRSs. These systems generated quantitative data in the form of matrices and frequencies. For further detail, refer to section 3.6.

Two sampling techniques were used in this research. During the exploratory research, purposive sampling was used, as specific forms of textual documents were needed to evaluate the functionality of the IRSs. During the explanatory research, performing the comparative evaluation between the two IRSs, systematic random sampling was used to create the final collection of 100 documents for the evaluation.

For this research, the unit of analysis was a query and the unit of observation, a document.

### **1.7.2 Data analysis**

Extensive data analysis was performed during the three pilot tests during the design, build, and test cycles. Supporting the first research question, IRS performance measurements were used to judge the effectiveness of the two IRSs. The results from the questionnaire provided user relevant and user non-relevant Boolean values converted to binary. The IRS generated data provided system retrieved and system not-retrieved frequency values converted to binary. Eight performance measurements were used to calculate the values of: Precision, Recall, Fallout, F-measure, Snobbery ratio, Specificity, Noise factor, and Accuracy.

### **1.7.3 Statistical analysis**

Statistical analysis was performed to test the five hypotheses. To test the first hypothesis, Precision (P), ranked average precision (AP) and mean average precision (MAP) were used, and a one-tailed student's t-test to test statistical significance was performed. To test the second hypothesis recall, ranked average recall (AR) and mean average recall (MAR) were used, and a one-tailed student's t-test to test statistical significance was performed. To test the third hypothesis, Specificity (S), ranked average specificity (AS) and mean average specificity (MAS) were used, and a one-tailed student's t-test to test statistical significance was performed.

To support the last two hypotheses, the Kappa coefficient was used to determine any differentiation between user and IRS judgements. Agreement measurements use a six-division range. IBM SPSS statistics version 25 (SPSS) was used to perform the statistical analyses for the one-tailed student's t-test and the Kappa coefficient.

## **1.8 Findings**

The eight headline findings evidenced in this research are:

- i) The hybrid indexing method maintains word ordinality and word proximity.

- ii) Hybridised indexing increases the effectiveness of retrieving relevant documents.
- iii) Hybridised indexing reduces the incorrect identification of relevant documents.
- iv) Hybridised indexing increases the quality in rejecting non-relevant documents.
- v) The judgments made by the hybrid indexing method and the user disagree.
- vi) The hybrid indexing method does not satisfy the information needs of the user.
- vii) The hybrid indexing method reduces reading time since it produces fewer non-relevant documents.
- viii) The design of the hybrid index solves the vocabulary mismatch problem of matching a query to a document.

### **1.9 Recommendations for further research**

Four key recommendations for further research became evident from this research, namely:

- i) The hybrid indexing method has the ability to match phrase-terms expressed within a query to those within a document exactly – This method should be used by those researchers and others who are in need of high precision, high specificity, and highly effective searching using IRSs.
- ii) Search engines should have options that the user can ‘set’ certain working parameters to achieve pure non-influenced search. For example, ignore parenthesis, ignore special characters, perform phrase-term exact matching, disallow synonyms, and remove weighting and ranking algorithms.
- iii) Judgments made between users disagree – The reasons why users make mistakes in judgements and how these mistakes can be avoided must be investigated and determined.
- iv) IRSs do not satisfy the information needs of the user – there is a need to better understand what it is that makes a user decide a document does not meet his/her information need.

### **1.10 Ethics**

As the researcher for this study, this author acknowledged that it was his responsibility to follow the Cape Peninsula University of Technology code of practice on ethical standards together with any relevant academic or professional guidelines in the conduct of the study. All computer software used in this research, Microsoft Access and Microsoft Visual Basic Access, is fully licenced. This researcher’s number-based ethics lie in the program code developed as well as how the data were treated in the development of three artefacts including the indexing methods. The intellectual

property of this research is shared 20% for CPUT and 80% for the author – this was arranged with the CPUT Technology Transfer Office (CPUT, 2019).

### **1.11 Delineation and limitations**

This study has four important limitations:

- i) A few documents in the collection were too large in length, limiting a user's ability to peruse them effectively.
- ii) The optical character recognition (OCR) software made a few errors in converting documents to text.
- iii) To ensure unwanted bias between IRSs, the two IRSs shared the same set of program code where the only differentiating factors were the use of query terms, apostrophes and the indexing method.
- iv) The document collections were closed computer based collections and not open Web based collections.

### **1.12 Contributions**

#### **1.12.1 Theoretical contributions**

From this research, the contribution to knowledge is the hybrid indexing method, which is simultaneously a theoretical contribution (the theoretical design, which combines and extends many concepts from the literature). This method takes the inverted index and combines these with the theoretical data retrieval property of exact matching, together with the key concept of the unique token identity number that maintains word ordinality and word proximity, and uses the measurement of phrase-term frequency.

#### **1.12.2 Methodological contributions**

Through the use of the hybrid indexing method, the methodological contributions provide more effective retrieval of special-interest documents. It allows for mismatching vocabulary using multiple synonymous phrase-terms, and uses the concept of exact phrase matching to increase precision, to reduce recall, and to increase the quality of specificity. This method allows for expanded phrase-term queries (used to better describe a user's information need) and exact phrase matching (to better match a query to a document) to improve precision, reduce the retrieval of non-relevant documents, and increase the quality of rejected non-relevant documents. By design, this research provides a partial solution to a practical problem by reducing the time required for the user to identify those documents relevant to his/her information need. This solution enables the user to perform multiple expanded phrase-term search queries and to retrieve more effectively the relevant documents within a shorter timeframe.

### **1.12.3 Practical contributions**

There are many practical contributions using the hybrid indexing method that apply to many industries. For further detail, refer to section 7.2.3.

- i) In postgraduate research, the hybrid indexing method can be used to identify documents that contain short or long text used in phrases or sentences.
- ii) A digital university library can use the hybrid indexing method to search for documents more effectively, thus eliminating the need for the use of keywords.
- iii) The motor industry can benefit from the hybrid indexing method when multi-word searches are used to find motor vehicles that are of a specific year, make, and model.
- iv) In the legal profession where many large libraries of legal documents exist, which, if digitised, can be accommodated by the hybrid indexing method. Searches for specific South African legal terms or the various Acts can be made effectively.
- v) In information systems implementation, metadata of a document are often used to describe and index a relationship with a document. It is often feasible to produce this metadata manually when document volumes are small but when they are large the metadata can be automated using techniques applied in information retrieval and the hybrid indexing method.

### **1.13 Summary**

Researchers battle to find documents effectively from their own personal collection pertaining to their information needs. Therefore, this study is not about Google and the Web but about a researcher's own collection of documents where the researcher needs to find documents that contain specific phrases. The problem statement for this study is as follows:

Challenges exist for information retrieval systems in handling mismatching vocabularies in queries and candidate source documents (Onal et al., 2018). As a result, these information retrieval systems may retrieve some documents that are non-relevant and miss some that are relevant (Van Gysel, 2017). This increases the time for research by forcing additional perusal of unsatisfactory results, and additional searches using alternative vocabularies (Liu et al., 2017). This renders information retrieval systems less effective than they could be, and inhibits productive research (Mitra & Awekar, 2017; Nguyen et al., 2018).

The research questions and hypotheses are as follows:

**RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

**H1<sub>0</sub>:** Hybridised indexing does not increase the effectiveness of retrieving relevant documents

**H2<sub>0</sub>:** Hybridised indexing does not reduce the incorrect identification of relevant documents

**H3<sub>0</sub>:** Hybridised indexing does not increase the quality in rejecting non-relevant documents

**H4<sub>0</sub>:** Judgments made by the hybrid indexing method and the user disagree

**H5<sub>0</sub>:** The hybrid indexing method does not satisfy the information needs of the user

**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

This research aimed to perform an exploratory study, using DSR by designing, building and then rigorously pilot testing a new IRS using a hybrid indexing method that maintains word ordinality and word proximity, and to provide the conceptual framework for this method. The second aim was to perform an explanatory study, via experimentation, to determine whether the hybrid indexing method solves the problem of mismatching vocabulary between a query and a document.

The headline findings are:

- i) The hybrid indexing method maintains word ordinality and word proximity.
- ii) The hybrid indexing method matches a phrase-term query to a document exactly.
- iii) Hybridised indexing increases the effectiveness of retrieving relevant documents.
- iv) Hybridised indexing reduces the incorrect identification of relevant documents.
- v) Hybridised indexing increases the quality in rejecting non-relevant documents.
- vi) The judgments made by the hybrid indexing method and the user disagree.
- vii) The hybrid indexing method does not satisfy the information needs of the user.
- viii) The hybrid index design solves the vocabulary mismatch problem of matching a query to a document.

The contribution to knowledge is the design of a novel hybrid indexing method. This author acknowledges that it is his responsibility to follow the CPUT code of practice on ethical standards.

### 1.14 The framework of the thesis

This thesis comprises two volumes: Volume I has seven chapters, and Volume II contains the appendices; the first three appendices contain the pilot tests. The two volumes are structured as indicated in Figure 1.3 below.

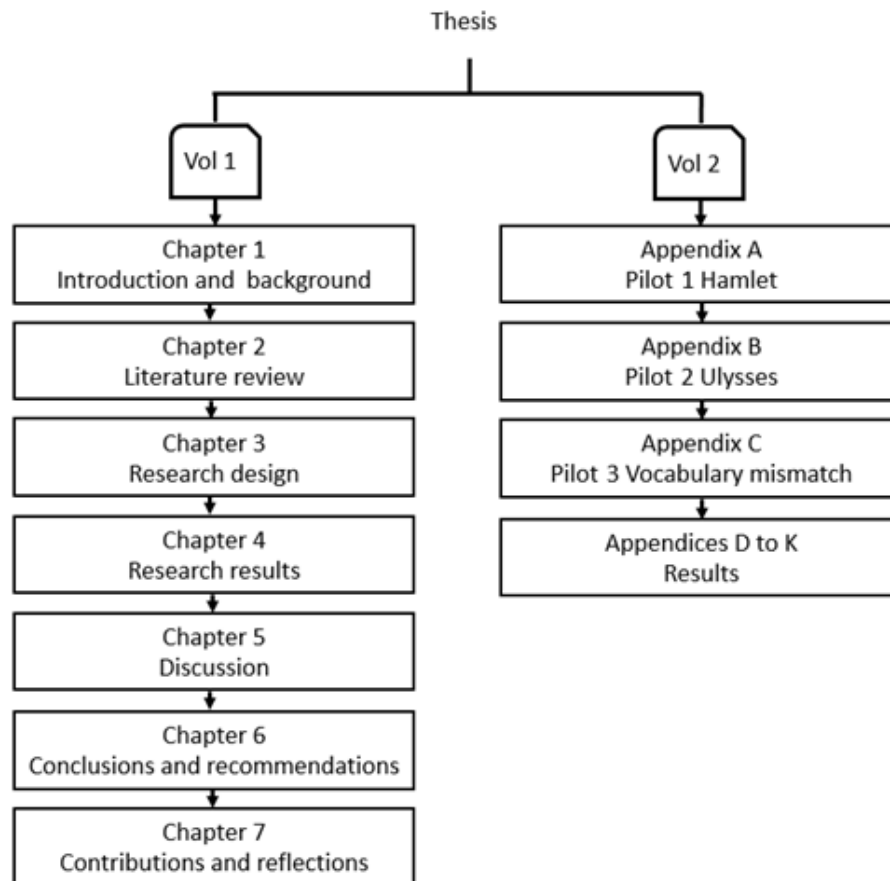


Figure 1.3: The framework of this thesis

#### VOLUME I

**Chapter One: Introduction and background** – the context of, and the approach to the study are provided in this chapter.

**Chapter Two: Literature review** – earlier work relevant to information retrieval systems and indexing methods is reviewed and a conceptual framework is developed.

**Chapter Three: Research design** – the choice of the design to the study is explained, and the research questions, hypotheses, data collection, and statistical analysis methods are discussed.

**Chapter Four: Research results** – the results of the study are brought together and discussed, and the results are presented.

**Chapter Five: Discussion** – the results are reviewed according to the aims, objectives, hypotheses and research questions of the study, and the findings are presented.

**Chapter Six: Conclusions and recommendations** – from the discussion and findings conclusions are drawn and recommendations are made for this research and further research.

**Chapter Seven: Contributions and reflections** – the theoretical, methodological, and practical contributions of this research are discussed followed by reflection, the assessment of design, and self-reflection.

## **VOLUME II**

**Appendices** – there are eleven appendices. The first three appendices describe and present the design, build, and test results for the three pilot tests based on DSR. The remaining eight appendices contain the expansive data using large tables (term-by-document matrices, phrase-term-by-document matrices, IRS judgements, user judgments, performance measurements) relevant to the results of this research.



## CHAPTER TWO: LITERATURE REVIEW

*“Inventions rarely come from people within an industry, but, instead come from people on the outside who aren't under the same limiting beliefs & habitual thinking that forms within any organisation or industry” – James Asher*

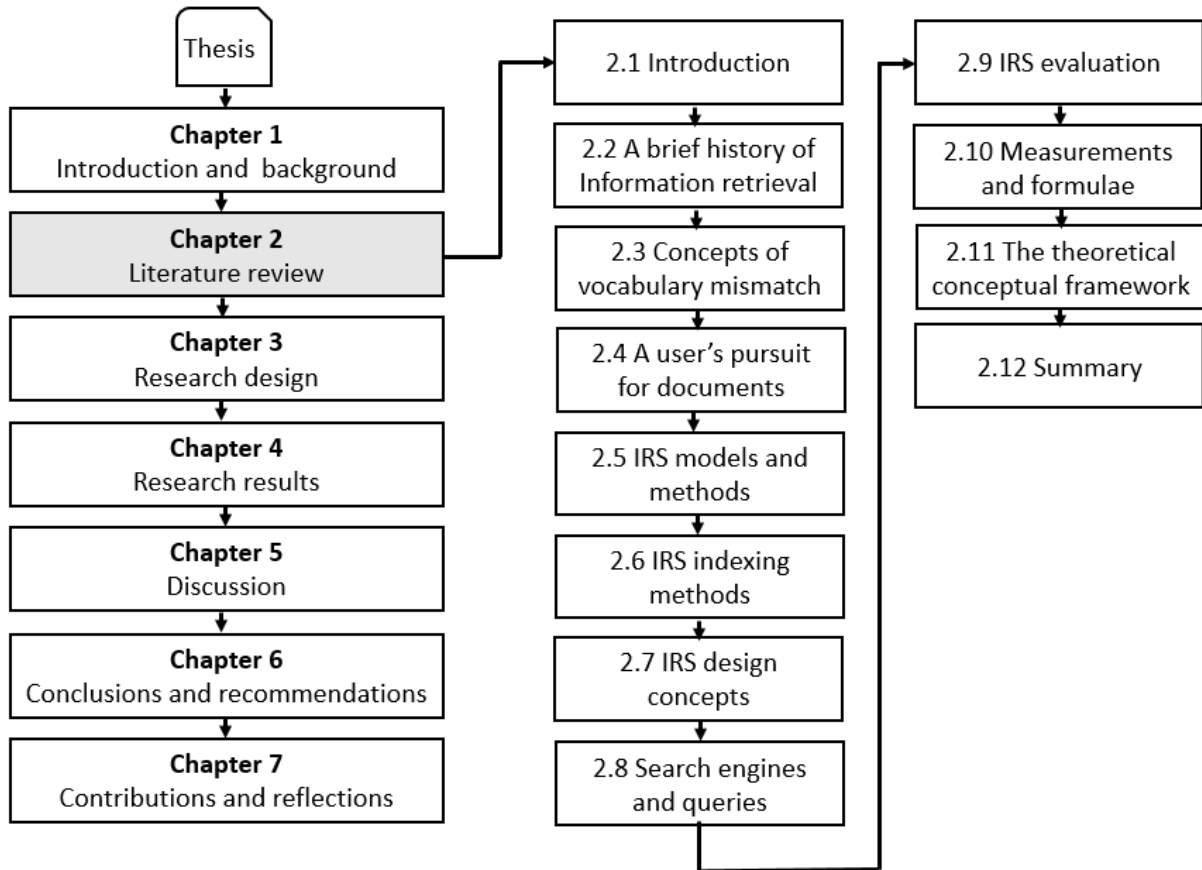


Figure 2.1: Schematic representation of Chapter Two

### 2.1 Introduction

The literature review focuses specifically on mismatching vocabulary of information retrieval indexing methods from a researcher's viewpoint. Furthermore, it explores and determines why challenges exist for information retrieval systems in handling mismatching vocabularies in queries and candidate source documents. The literature strategy determines the literature gap in order to help understand why the research problem exists and trying to determine what is actually missing that has created the root cause of the problem. For the benefit of the reader, the two research questions and the five null hypotheses are re-stated:

**RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

**H1<sub>0</sub>:** Hybridised indexing does not increase the effectiveness of retrieving relevant documents

**H2<sub>0</sub>:** Hybridised indexing does not reduce the incorrect identification of relevant documents

**H3<sub>0</sub>:** Hybridised indexing does not increase the quality in rejecting non-relevant documents

**H4<sub>0</sub>:** Judgments made by the hybrid indexing method and the user disagree

**H5<sub>0</sub>:** The hybrid indexing method does not satisfy the information needs of the user

**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

The first aim of this research was to perform an exploratory study, using DSR, by designing, building and then rigorously pilot testing a new IRS using a hybrid indexing method that maintains phrase-term word ordinality and word proximity. The second aim was to perform an explanatory study, via experimentation, by comparing the effectiveness of the hybrid indexing method with that of the traditional inverted indexing method. Thereafter to prove statistically that the hybrid indexing method increases the effectiveness of retrieving only those documents that are judged relevant by the user; reduces errors in incorrect identification of user judged relevant documents; and increases the rejection quality of user non-relevant documents. Finally, to determine whether the problem of mismatching vocabulary can be solved and whether the information needs of the user can be satisfied.

From the problem statement, research questions, hypotheses and research aims, key concepts were selected and used to search for relevant literature and these are: indexing method, information retrieval system, information gathering, information need, search engine, vocabulary mismatch, query, user judgment, performance measurement, precision, recall, and F-measure, among others. To enrich the literature and to ensure an expansive review, these key concepts were used to search the literature using university specific online databases, including EBSCOhost<sup>3</sup>, ProQuest<sup>4</sup>, Emerald<sup>5</sup>, and Scopus<sup>6</sup> from CPUT and UCT, Google search and Google Scholar for Web based pdf documents relating to journal articles, conference papers

---

<sup>3</sup> <https://www.ebsco.com/products/ebscohost-platform>

<sup>4</sup> <https://www.proquest.com/libraries/academic/databases/>

<sup>5</sup> <https://www.emeraldinsight.com/>

<sup>6</sup> [https:// www.scopus.com/home.uri](https://www.scopus.com/home.uri)

and theses, and finally the Web sites of Takealot<sup>7</sup> and Amazon<sup>8</sup> for the purchasing of key concept specific books.

The literature review is now presented, beginning with a brief history of information retrieval, an introduction to the concepts of vocabulary mismatch, followed by a report of a user's pursuit for relevant documents. Thereafter the focus is on the IRS and the models and methods described in the literature together with indexing methods and design concepts. Search engines and the queries they utilise are explored, followed by the methods used to evaluate IRSs together with the measurement instruments and formulae used for evaluation. Finally, the theoretical conceptual framework derived from the literature is presented with a summary.

## **2.2 A brief history of information retrieval**

Information retrieval is a process of retrieving information from a document with an element of uncertainty relating to chance, where information retrieved is judged relevant or non-relevant. This element of chance is described theoretically in the early work by Bayes (1763) entitled, '*An essay towards solving a problem in the doctrine of chances*', where the author introduced a set of mathematical probabilities to help solve problems relating to chance, now known as Bayes theorem. Nearly two centuries later, Gross and Gross (1927) introduced citation analysis and were the first to use citations to evaluate scientific journal importance, based on the theories of chance. Gross and Gross (1927) tabulated the 1926 Journal of the American Chemical Society references, in the form of a document-by-citation matrix, and ranked the importance of cited journals to chemistry students (Pinski & Narin, 1976). Zipf (1935, 1965), in his work entitled, '*The Psycho-biology of language: an introduction to dynamic philology*', discovered that the distribution of words in the English language takes the form of an harmonic series where, on average, the first most frequently used word in a document collection would occur every ten words, the second most frequently used word would occur every 20 words, etc. Many years later, in the year that World War II (WW II) ended, Bush (1945) introduced the idea of automatic access to huge volumes of stored information and knowledge, and emphasised the need of indexing for the successful selection and retrieval of information. Bush (1945) predicted a mechanised device called a *memex* that in the future would store books and documents, and when consulted, would retrieve information quickly.

---

<sup>7</sup> <https://www.takealot.com/>

<sup>8</sup> <https://www.amazon.com/>

Shortly after WW II, Gray (1947) discussed a patent in his work entitled, '*Pulse code communication*', where encoded numbers allowed adjacent numbers to have a single digit differing by 1, and introduced the concept of a *tuple* or *n-tuple* – a finite ordered list of elements. The n-tuple is now referred to as an n-gram where  $n = 1, 2, 3, \text{ etc.}$ , denoting the number of words in a phrase (Ha et al., 2002) and is perceived to be an early form of the *k*-word proximity search that is discussed later in this chapter. Zipf (1949) described the principle of least effort (PLE) now known as Zipf's law. What Zipf (1949) had done, was to discover through a manual analysis of the book '*Ulysses*' by Joyce (1932) that the frequencies of word types was a fraction of the total number of word tokens<sup>9</sup>. The result was that 29,899 distinct word terms<sup>10</sup> were associated with 260,430 word tokens (Zipf, 1949; Ha et al., 2002) and produced a formula  $f = \frac{k}{r}$  where *f* is the frequency of a word in the data collection, *r* is the rank, and *k* is a constant for the document collection. A few years later, Mooers (1950; 1951) introduced the phrase *descriptor* as the origin of the query term, coined the phrase *information retrieval*, and introduced Zatocoding, now known as *hashing*. Mandelbrot (1953) stated that language has three elements: i) the structure of the language; ii) the way in which information is coded in the brain; and iii) the economical criterion of matching that links point 1 to 2. In addition, Mandelbrot (1953) suggested modifications to Zipf's law and introduced the formula  $f = \frac{k}{(r+\alpha)^\beta}$  where  $\alpha$  and  $\beta$  are constants for the data collection under analysis (Ha et al., 2002). Luhn (1953) theorised a new method for recording and searching for information that provided responses in all cases, not only the relevant cases, paving the way for the use of the terms relevant and non-relevant, the two-class classifier<sup>11</sup> known as the 2x2 contingency table, and the F-measure mathematical formula based on Precision and Recall. Harris's (1954) linguistic work entitled, '*Distributional structure*', introduced the '*bag of words*' concept that referred to a collection of words randomly dropped into a bag, without structure and without order. In the same year, Perry et al. (1954) theorised information retrieval automation and argued the advantages that automated information retrieval could provide. In the inspiring work of Garfield (1955) on the science of citation analysis, the author laid the foundation for the ideas and concepts for the citation ranking method, a concept used in the Google search engine today

---

<sup>9</sup> A token is defined as a chunk of text separated by spaces contained within a document which can take the format of a word within a language or a term – in an IRS these tokens are populated within an index and are made available for searching purposes

<sup>10</sup> A term is defined as a chunk of text separated by spaces contained within a document which can take the format of a word within a language or a term – in an IRS a term is expressed within a query which is presented to the search engine

<sup>11</sup> A classifier is a function that takes objects and assigns them to one or more distinct classes

(Brin & Page, 1998). Kent et al. (1955) and Cleverdon (1956) introduced the concepts of *recall* and *precision* together with their mathematical formulae, and Simon (1955, 1996) described a derivation of Zipf's law (Zipf, 1949) in his work entitled, '*On a class of skew distribution functions*', based on empirically derived distribution functions (Ha et al., 2002).

During WW II, huge volumes of technical reports relating to engineering and aeronautics were written and after WW II, indexing systems to retrieve these documents needed upgrading due to their sheer volume. Cleverdon (1956) was a librarian working at the Cranfield College of Aeronautics at that time and had a passionate interest in indexing systems. Because of Cleverdon's work in aeronautics, he was requested to work with the North Atlantic Treaty Organisation (NATO) where he was introduced to the Uniterm system. The Uniterm system of indexing was invented by Taube (1956), a government librarian, when he discovered that 40,000 subject headings in a card catalogue contained only 7,000 distinct words (Cleverdon, 1991). In the work of Luhn (1957) entitled, '*A statistical approach to mechanized encoding and searching of literary information*', Luhn argued that because of users choices, word combinations, linguistic meaning and different levels of specificity, literature searching by machines would still present challenges. However, three years later, Cleverdon (1960) argued that the requirements for the design of these machines to perform as IRSs would eventually be possible. Building on Precision and Recall by Kent et al. (1955), Cohen (1960) introduced a measurement for relevancy known as the Kappa coefficient that measured agreements between judges. In the work of Levenshtein (1965) entitled '*Binary codes capable of correcting deletions, insertions, and reversals*', Levenshtein introduced a measurement between two strings now known as the Levenshtein distance. The Levenshtein distance, or edit distance, described by Gusfield (1997), is a similarity measurement between two character strings (referred to as '*tokens*' in this research). Rocchio (1965) and Rocchio and Salton (1965) presented relevance feedback for IRSs described as an iterative process whereby a user fine tunes queries in an attempt to retrieve relevant documents, and these authors introduced a measure now known as the Rocchio algorithm. Adding to the **ratio concept of recall** – also referred to as *hit rate* or *sensitivity* (Cleverdon & Keen, 1966, citing Western Reserve University), and the **ratio concept of precision** – also referred to as *relevance ratio*, *pertinency*<sup>12</sup> *factor*, *acceptance rate* (Cleverdon & Keen, 1966) and *noise factor* (Cleverdon & Keen, 1966 citing Perry), Cleverdon and Keen (1966, citing Fairthorne) introduced the *snobbery*

---

<sup>12</sup> To engage in noisy celebration

*ratio*, *fallout ratio* (Cleverdon & Keen, 1966) and *specificity*<sup>13</sup> (Cleverdon & Keen, 1966 citing Western Reserve University) together with their mathematical formulae. About this time Henderson (1967:119), referring to a physical file, states: “Specificity takes into account one of the vital parameters in a retrieval system ... size of file”. Six years later, Spärck Jones (1972) introduced the theory and use of inverse document frequency (*idf<sub>i</sub>*), a weighted formulation, rather than the traditionally used non-weighted document frequency (*df<sub>i</sub>*), and shortly afterwards Akaike (1974) introduced the Akaike information criterion, a theoretical quality measure for statistical model data sets trading off distortion against the model’s complexity.

In 1975, Salton, Wong and Yang (1975) presented the vector space model for automatic indexing. This model has now become one of the oldest and most extensively studied models for using theories from linear algebra (Manning et al., 2008). A year later, Harris (1976) presented a formal theory of language structure encompassing three fundamental relations: ordered-entry discourses, reduction, and the entry and reduction system. Pinski and Narin (1976) built on the ideas and concepts of citation analysis, ranking, and the work of Gross and Gross (1927), and introduced citation influence methodology for scientific publications. In linguistics, Harris (1979) argued that language can be segmented into successive sentences, and that each sentence is a representation of a sequence of words, thus introducing the concepts of stems with affixes, morphemes<sup>14</sup>, and phonemes<sup>15</sup>.

Van Rijsbergen (1979) referred to the complement of the F-measure stated as  $E = 1 - F$  and described the distinguishing properties differentiating data retrieval from information retrieval. In 1980, Porter (1980) announced the suffix stripping algorithm, an automatic method for removing suffixes from English words thus assisting information retrieval processes. Robertson (1981) argued that no definitive method for evaluating IRSs existed, and further argued that when new systems were developed, and the inadequacies of the old systems were revealed, these inadequacies would just be replaced by new challenges arising from the newer system. Blair and Maron (1985) added to the list of challenges and argued that in information retrieval, to predict the exact words (for indexing) and their combinations, and the phrases that users make use of, is unbelievably difficult. A few years later, Salton and Buckley (1988) confirmed the use of effective term weighting systems as used earlier by Spärck Jones (1972) and concluded in their work that these produced

---

<sup>13</sup> The quality of being specific

<sup>14</sup> A meaningful morphological (the study of words) unit of a language that cannot be further divided

<sup>15</sup> A distinct unit of sound in a specified language distinguishing one word from another

superior results for text indexing systems. A year later, Berners-Lee (1989) invented the World Wide Web (Web) at the European Organisation for Nuclear Research in Geneva and within a few years the Web became the world's largest document collection.

Cleverdon (1991) reflected back on the significance of the Cranfield tests on index languages, while Garfield (1997) looked at the work of Mooers, all in the 1950s. Brin and Page (1998) developed the PageRank citation ranking measurement and thereafter founded Google Incorporated. In essence, what Berners-Lee (1989) provided theoretically and practically by inventing the Web was the method to link one document to another, while Brin and Page (1998) helped the user to search and find that document link. Improvement to the vector space model developed but relevancy still remained a challenge, therefore Berry, Drmac and Jessup (1999) pointed out the need for a cosine similarity threshold (a selected tolerance value) to be used when judging document relevancy. Clarke, Cormack and Tudhope (2000) suggested improvements through their short query ranking measure, called *cover density*, that expanded coordination level ranking through the measurement of term co-occurrence. The recent previous head of search for Google Incorporated, Singhal (2001), reflected on the information retrieval past and recalled the four most used theoretical information retrieval models used in research, namely the Boolean retrieval model, the vector space model, the probabilistic model(s), and the inference network model.

In more recent work there are many interesting models relating to IRSs. Referring to the vector space model (Castells, Fernandez & Vallet, 2007; Langville & Meyer, 2007) based on the early work of Salton and Buckley (1983), Castells et al. (2007) and Binkley and Lawrie (2015) describe the vector space model as a model, whereas Langville and Meyer (2007) and Chew et al. (2011) refer to it as a method. In this research, the vector space is referred to as a model; secondly, the *n-tuple* Gray code in the work of Losee (2006) is based on the work of Gray (1947); and thirdly, the three theoretically developed semantic search models that are introduced in the work of Koopman (2014) (though not utilising traditional term-based but rather concept-based queries), are: i) the bag of concepts model, fundamentally represented utilising concepts of healthcare ontology; ii) the Graph-based Concept Weighting model, that introduces a novel weighting function capturing concept dependence and importance; and iii) the Graph INference model, developed through the integration of ontologies and statistical information retrieval methods. The results reveal that additional previously undiscovered documents can be retrieved, thus expanding the corpora

through concept-based queries rather than by utilising term based queries. From this work, the evidence represents a leap forward in the integration of ontology (structured domain knowledge) and term based information retrieval methods. The author concludes that the semantic search model evidences how the results from traditional information retrieval corpora<sup>16</sup> methods can underestimate effectiveness of such methods and suggests that new methods are explored.

One of the major solutions to many of the information retrieval challenges is indexing. In Shoaf's (2013) review of the contributions to the theory of indexing and information retrieval by Cleverdon (1960) and Cleverdon and Keen (1966), the author summarises interesting facts: i) the subject knowledge of the staff participating in the search experiments is directly related to the best results retrieved, as the knowledgeable staff member would put much effort into the search query, because the exact meaning of the question was better understood; ii) Cleverdon discovered that single word term indexing languages were superior to other indexing languages; iii) natural language indexing gave reasonable performance; iv) as the searched total number of documents generated increased the usefulness of these documents decreased; v) recall and precision improvements were based on the knowledge and skill of the searcher and the familiarity of the indexing system. As Shoaf (2013) emphasises, this was all evidenced at a time of computer infancy. In the work of Cleverdon (1956), the author specifies the proposed indexing parameters to be used during the experiments, and in summary, these are 20,000 documents that were required to be indexed by five different systems within a two-year period by three people (the indexers). The important control measures revealed were the time taken to index each document and the identity of each indexer indexing each document, as humans naturally judge things differently (Cleverdon, 1956). The discussion around the design and challenges of information retrieval is now put to one side in order to explore what information retrieval actually is in modern day terminology.

### **2.3 Concepts of vocabulary mismatch**

Vocabulary mismatch is a phenomenon whereby multiple words in a phrase used to describe something in the past change over time in order to describe the same thing (Shekarpour et al., 2017). When these phrases are expressed within a search query, a mismatch occurs between the query and the document (Onal et al., 2018). Vocabulary mismatch thus impacts the effectiveness of text based IRSs as the words

---

<sup>16</sup> Multiple large sets of stored texts



within the phrases expressed as queries are not accurately matched to the words within the text of documents within a collection (Nguyen et al., 2018).

In attempting to solve the problem of mismatching vocabulary there are a few challenges. Using the correct words in a query to describe what one is looking for is the first challenge. The second challenge is that the user needs to use words in the query that actually exist in the document. If the words are chosen incorrectly, a mismatch occurs between the query and the document. This problem of mismatching vocabulary was first hinted at by Sticht, Beck and Hauke (1974) and more formally presented in IRSs by Furnas et al. (1987) as the vocabulary problem. Since Furnas et al. (1987) coined the phrase *vocabulary problem*, a large volume of work has been done. This vocabulary problem<sup>17</sup> was discussed by Turtle and Croft (1991), pioneers in IRS theory and by Egoli, Markovitch and Gabrilovich (2000). Turtle and Croft (1991) argued that the reason for poor retrieval of documents was because of poor matches between the vocabularies used expressed within a query and the vocabulary used within documents. However, Egoli et al. (2000) discussed the limitations of indexing and suggested that the indexing design was the problem. Egoli et al. (2000) argued that since keywords were introduced years ago and that indexing methods had not changed over time, keywords had become *noisier*<sup>18</sup> (imprecise) especially to a non-expert user, thus creating the vocabulary mismatch problem.

The phrase *vocabulary mismatch* is itself a vocabulary mismatch problem as various authors have described vocabulary mismatch in differing ways. For example, the phrase *vocabulary problem* has evolved over time into *vocabulary gap* (IJzereef, Kamps & De Rijke, 2005), *vocabulary mismatch* (Min et al., 2010), *term mismatch* (Sirres et al., 2018) and *semantic gap* (Nguyen et al., 2018; Koopman & Zuccon, 2019). Antonyms have been used to describe the opposite, for example, *vocabulary agreement* (Chaparro, Florez, & Marcus, 2016) and *vocabulary normalisation* (Binkley & Lawrie, 2015).

Vocabulary problem, vocabulary gap, vocabulary mismatch, term mismatch, and semantic gap are five bi-word synonymous phrases that have similar meaning. The phenomenon of vocabulary mismatch occurs in IRSs when the words within a search query mismatch the words within a document (Onal et al., 2018). A few authors refer to this phenomenon as vocabulary gap, where a gap is created between the search

---

<sup>17</sup> The phrases 'vocabulary problem', 'vocabulary mismatch' and 'vocabulary mismatch problem' are used interchangeably – more recent work is discussed shortly

<sup>18</sup> Refers to noise factor, which is the measure of degradation within a system

queries and the documents when different words are used to describe the same concept (Liu et al., 2017; Van Gysel, 2017).

The vocabulary mismatch problem is twofold: firstly, there is a user usage problem whereby a mismatch occurs between the words expressed in a query and those words that exist in the text of a document. This problem is compounded since the use of the words change over time. Secondly, there is a design problem, whereby the index containing the words in the search query mismatches the index containing the words from the document. Referring to the 'usage problem' and to the year 2016, Onal et al. (2018) state that in text information retrieval, relevant IRS research has been concentrated on the long-standing problem of vocabulary mismatch. Onal et al. (2018) confirm vocabulary mismatch remains a problem and define vocabulary mismatch as a phenomenon where the vocabulary used in relevant documents and the vocabulary of the person searching for the document may differ.

A number of attempts have been made to solve the first vocabulary mismatch problem. In the work of Koopman et al. (2016), the authors present a graph inference retrieval model for complex queries with the aim of improving information retrieval in the biomedical domain. In the work of Goeuriot et al. (2016), the authors recommend that inference:

- i) can improve retrieval when using complex queries, and
- ii) can create a more effective IRS by retrieving documents additional to those retrieved using traditional approaches.

Many authors have approached the vocabulary mismatch problem in different ways and a few of these are now discussed. He and Ounis (2009) investigated the ineffectiveness of query expansion and that this ineffectiveness is based on IRSs retrieving too many 'non-relevant' documents because:

- i) too many expansion terms are expressed in the query, and
- ii) although IRSs retrieve documents, they are often irrelevant to the information need thus the use of expansion terms is deemed problematic.

He and Ounis (2009) conclude that query expansion does not always provide an increase in IRS effectiveness.

Hanbury et al. (2014) reviewed seven papers pertaining to intellectual property within the legal domain that acknowledged certain factors that affected those users searching for patents. The two primary factors evidenced were the multimodal and multilingual format of the data being searched and the need for an appropriate

strategy (which included query expansion) for searching the data. Of the seven papers, three suggested using query expansion, one paper considered the use of multiple query expressions to improve patent search retrieval results, and the final three papers covered patents for multilingual retrieval of patents, text categorisation, and image retrieval.

In the work of Pal et al. (2015), the authors evidenced methods using query expansion that could automatically classify a given query into one or more pre-defined categories. Pal et al. (2015) hypothesised that overall IRS performance would improve if specifically personalised query expansion techniques were applied to a given query, rather than applying general query expansion techniques to all the queries. In the conclusion of their work, Pal et al. (2015) propose the taxonomy of query classes and recommend that from a query expansion perspective, query categorisation should be considered.

Soldaini et al. (2016) investigated a utility to bridge the vocabulary gap between the non-expert and the expert, and therefore aimed their work at improving medical information retrieval, in order to assist the medically uninformed to find medically phrased information through the use of query clarification – a form of query expansion where multiple words in a search are used to better express an information need. Soldaini et al. (2016) argue that the language gap is one of the main reasons why IRSs fail.

A few authors of recent research (Van Gysel, 2017; Onal et al., 2018) have focused on improving the first vocabulary mismatch problem through a better understanding of the user's intent in the search query. Onal et al. (2018:17) describe the concept of "query understanding" and explain that some publications are focused on distributed representations of queries and the use of similar queries can better express the intent of the user.

Van Gysel (2017) investigated the formulation of queries and introduced a query formulation model. Query formulations are ways in which queries can be expressed, used by experts and non-experts, in an attempt to help solve the vocabulary mismatch problem. In the conclusion of his work, Van Gysel (2017) confirms the effect of terms used in queries and the existence of the vocabulary mismatch problem and argues that high relevance does not mean a high matching degree at the term level, and vice versa, as those documents that match zero query terms could still remain relevant.

In their work, Onal et al. (2018) used three distinguishing tasks to understand queries better: query suggestion, where the IRS pre-empts a query and makes a suggestion; query auto completion, where the IRS typically suggests queries used in previous searches; and query classification, where search queries are assigned to one or more predefined categories.

The concepts and designs of indexing have evolved over many years (Taube, 1956; Spärck Jones, 1972; Brin & Page, 1998; Panigrahi & Gollapudi, 2013; Croft, Metzler & Strohman, 2015). Indexing methods include the phrase index (Ha et al., 2002); the next word index (Williams, Zobel & Bahle, 2004); the tiered index (Panigrahi & Gollapudi, 2013); the inverted index (Croft et al., 2015); the positional inverted index (Procházka & Holub, 2017); and others. Further details of these indexing methods are discussed in section 2.6.

However, research focus appears to have shifted away from index design, to query design to solve the vocabulary mismatch problem. For example, and to explain this shift, concepts and mathematical formulae have been introduced to try to influence a query's effectiveness:

- (i) term frequency (*tf*) – a measure for the number of times a term occurs in a document (Kang et al., 2015);
- (ii) collection frequency (*cf*) – a measure of the number of times a token occurs in a document collection (Perry et al., 1954; Van Gysel, De Rijke & Kanoulas, 2017);
- (iii) inverse document frequency (*idf*) – the inverted measure of document frequency that attempts to suppress the effect of frequently occurring terms (normally ignored) referred to as *stop words* (Spärck Jones, 1972);
- (iv) the vector space model where queries and documents are represented by vectors and an attempt is made to match the vectors (Salton et al., 1975); and
- (v) query expansion – a method that uses multiple terms expressed in a single query in an attempt to improve document retrieval performance (Zhao, 2012).

If a user uses words to describe a concept in its original form, and executes a search query, the best an IRS can do is retrieve the documents that contain those specific words<sup>19</sup> existing as text in the document, referred to as “hits”. This is referred to as Recall<sup>20</sup>, a measurement using a mathematical formula comparing user relevant

---

<sup>19</sup> Expanded queries can be used to overcome inference and the challenge of synonyms and homonyms and other words within the English language that exist in documents and is one way of attempting to solve the vocabulary mismatch problem

<sup>20</sup> The word Recall represents a formula and is therefore capitalised

documents to IRS retrieved and not retrieved documents. If an IRS retrieves all documents within a collection pertaining to the search query, Recall will reach 100%. The second formula is Precision<sup>21</sup>, which measures the matching quality of the words chosen, by comparing user relevant documents to IRS retrieved documents. If chosen correctly. Precision can theoretically reach 100% representing a perfect match of words within the query best describing the user's information need. The third formula is F-measure<sup>22</sup> which uses the values of Precision and Recall to measure the overall effectiveness of the IRS, but when Precision increases, Recall decreases and vice versa (Croft et al., 2015). Note that the inverse relationship between Recall and Precision refers to the overall performance of an IRS. Therefore, the inverse relationship does not necessarily apply to every single query presented to the IRS.

#### **2.4 A user's pursuit for documents**

In research, a user often has the requirement to seek documents pertaining to a particular subject. This involves the use of language, reading the document and deciding, a judgement, whether a document is relevant to the subject or not. The subject in IRS theory is referred to as the user's information need. According to Case (2002:5): "An information need is a recognition that your knowledge is inadequate to satisfy a goal that you have". However, rather than the need, other authors refer to the seeking of information (Kuhlthau, 1991), the human use of information (Dervin, 1992) and the the behaviour of the human (Wilson, 2000).

This section begins with the theory of language structure, then moves on to the information need, the relevance of documents, and the differentiating factors that exist between traditional information technology data retrieval and text information retrieval.

In the work of Harris (1976), the author presents a formal theory of language structure: the structure and information of sentences. This theory encompasses three fundamental relations: (1) ordered-entry discourses – the order of words that make up a sentence (or text); (2) reduction – the act of reducing a sentence into smaller quantities or words; and (3) the entry and reduction system (string or term rewriting systems) – when reductions are applied to ordered-entry discourses they effectively characterise all the sentences of a natural language. In summary, when words are applied within a specific entry order, a sentence is created using a natural language. This forms the basis of the bag of words model in information retrieval where, in earlier work of Harris (1954:11), the author states: "language is not merely a bag of words",

---

<sup>21</sup> The word Precision represents a formula and is therefore capitalised

<sup>22</sup> The word F-measure represents a formula and is therefore capitalised

where a bag of words includes a large variety of words (multiplicity) that excludes grammar and word order, while a sentence in a natural language does not. The concept of a “bag of words” is used to describe a subset of characteristics for a natural language as suggested in the earlier work of Harris (1954:11).

In the work of Gross and Gross (1927), the authors describe a *local need* of students wishing to gain access to scientific journals within a library. At that time in 1927 the accelerator for this need was pressure from students wanting to be become adequately qualified professional people and so wished to pursue postgraduate studies, and in particular, doctorates. To be able to do this, the students not only had to have access to the required journals but also a method to access the information pertaining to their desirability. The term *desirability* used by Gross and Gross (1927) is explained as once the need increases, then one particular journal may become more desirable than another. The terms *local need* and *desirability* used by Gross and Gross (1927) can be directly related to what we call an *information need* (Singhal, 2001) and *relevance* (Van Rijsbergen, 1979) in information retrieval theory. To support this need by the students and to measure desirability, Gross and Gross (1927) performed a citation analysis of the journals used by the students and the citations contained within them.

Table 2.1 presents the results from Gross and Gross (1927:2). The table resembles a document-by-citation year matrix, an early version of what is now referred to as a term-by-document matrix. What is interesting is that Gross and Gross (1927) list the documents as the rows and the years as the columns; similarly, today the traditional term-by-document matrix list the documents as rows and the terms as columns (this will be discussed later in this chapter). Additional features of their Table 2.1 include a frequency count of the citation occurrences within year ranges, together with a ranking system (Pinski & Narin, 1976; Brin & Page, 1998) sorted in descending order of the total number of citation occurrences (Garfield, 1972).

Although Gross and Gross (1927) within their table (Table 2.1) found 3,633 citations in 247 journals, for convenience they only list the top 28 journals (this can be related to the concept of the top-*k* documents (Manning et al., 2008) used in information retrieval today which is discussed later in this chapter). Gross and Gross (1927) argue that the number of citations is not the only criterion of desirability, as some journals with fewer references might be more desirable than others, possibly due to a higher quality. These theories of Gross and Gross (1927) can be related to what we use in information retrieval today, for example, citation occurrences can be related directly

to term frequencies and desirable journals of a high quality (Akaike, 1974), as document relevance (Van Rijsbergen, 1979).

Table 2.1: An early document-by-citation year matrix (Redrawn from Gross & Gross, 1927:2)

	Total	1921-1925	1916-1920	1911-1915	1906-1910	1901-1905	1896-1900	1891-1895	1886-1890	1881-1885	1876-1880	1871-1875
Ber.	686	78	30	67	115	79	64	60	56	53	44	33
J. Chem. Soc.	390	122	37	60	45	47	21	20	5	2	1	...
Ann.	278	26	8	37	33	23	22	21	19	18	13	...
Z. physik. Chem.	191	53	6	21	29	19	28	16	6	...	...	...
Compt. rend.	126	26	3	23	15	23	15	21	7	9	8	...
J. Phys. Chem.	93	42	13	13	5	1	1	...	...	...	...	...
Ann. Physik	93	18	4	28	13	6	0	0	6	5	2	...
J. Biol. Chem.	80	41	16	14	7	...	...	...	...	...	...	...
Am. Chem. J.	70	...	...	9	21	20	14	8	4	2	1	...
Z. anorg. Chem.	68	21	11	5	8	11	6	2	...	...	...	...
Ann. Chim.	68	5	0	6	9	7	3	5	1	8	4	2
Bull. Soc. Chim.	60	16	3	4	7	10	4	4	3	4	2	1
Proc. Roy. Soc.	55	30	5	4	8	5	1	0	1	...	...	...
J. Ind. Eng. Chem.	53	33	10	5	1	...	...	...	...	...	...	...
Z. Phys.	51	41	5	...	...	...	...	...	...	...	...	...
Monatsch.	51	2	1	21	5	9	3	2	5	3	...	...
J. prakt. Chem.	50	6	1	2	2	6	3	12	6	6	2	2
Phil. Mag.	49	17	14	4	2	3	3	1	1	0	0	1
Gazz. chim. ital.	44	10	6	2	6	4	8	4	3	0	1	...
Phys. Rev.	44	23	8	3	5	4	...	...	...	...	...	...
Physik. Zeit.	41	26	0	7	3	...	...	...	...	...	...	...
Z. Elektrochem.	37	11	13	4	4	4	1	...	...	...	...	...
Biochem. Z.	37	18	2	9	10	...	...	...	...	...	...	...
Rec. trav. chim.	36	14	5	2	2	2	5	4	1	1	...	...
SCIENCE	27	22	3	...	...	...	...	...	...	...	...	...
Trans. Far. Soc.	24	18	0	1	0	1	...	...	...	...	...	...
Proc. Nat'l Acad.	22	19	0	...	...	...	...	...	...	...	...	...
Nature	21	13	5	1	...	...	...	...	...	...	...	...

In the work of Garfield (1972) entitled, '*Citation analysis as a tool in journal evaluation*', the author introduces the term *impact factor* as a citation based size independent measure and concludes that, for science policy studies, journals can be ranked by frequency and impact. Garfield (1972:2) exemplifies this with journal citation frequencies in Table 2.2, which illustrates the occurrences of each journal cited during the last quarter of 1969 including the distribution by publication year of the particular issues cited. The list was compiled from more than 20,000 journals, books, reports, theses, and other documents cited during the last quarter of 1969 in journals covered by the Science Citation Index (SCI) (Garfield, 1972; 2007).

Table 2.2: Journal citation frequencies (Redrawn from Garfield, 1972:2)

ITEM NO	CITED JOURNAL	TOTAL	NUMBER OF TIMES CITED										
			1969	1968	1967	1966	1965	1964	1963	1962	1961	1960	REST
00243	ACTA PATH JAP	36	1	3	3	4	6	7	3	.	.	3	6
00244	ACTA PATH MICROBIOL	736	29	69	87	59	56	59	44	48	20	31	234
00909	AM J ANAT	637	7	27	37	56	32	41	15	26	15	21	360
00910	AM J BOT	1171	13	74	87	68	73	66	57	57	47	49	580
00911	AM J CANCER	103	.	.	.	.	.	.	.	.	.	.	103
00912	AM J CARDIOL	1238	73	201	199	247	134	70	78	66	53	73	44
03591	CAN J SOIL SCI	33	.	2	1	4	1	2	6	6	3	2	6
03592	CAN J SURG	61	2	6	4	3	13	11	3	3	5	1	10
03593	CAN J TECH	3	.	.	.	.	.	.	.	.	.	.	3
03594	CAN J ZOOL	356	46	38	40	28	24	20	19	29	17	22	73
08990	ISRAEL J AGR RES	29	1	1	7	.	1	8	7	2	1	.	1
08991	ISRAEL J AGR RES	16	.	.	3	5	4	3	1	.	.	.	.
08992	ISRAEL J CHEM	91	14	25	18	10	11	6	7	.	.	.	.
09651	J INVEST DERM	695	24	78	81	69	65	46	30	31	34	22	215
09652	J IOWA MED SOC	13	.	.	.	5	.	.	1	2	.	.	5
09653	J IRISH MED ASS	16	1	3	4	3	3	.	.	.	.	.	2
13390	P CALIF ACAD SCI	18	.	.	1	4	3	.	1	.	.	.	9
13391	P CAMBRIDGE PHIL SOC	389	8	22	23	11	12	9	13	11	17	3	260
19755	Z ANGEW CHEM	47	.	.	1	.	.	1	.	.	1	1	43
19756	Z ANGEW ENT	35	.	.	1	1	4	2	4	1	1	5	16
19757	Z ANGEW GEOL	49	2	7	5	8	5	5	4	4	2	1	6
19758	Z ANGEW MATH	10	1	.	1	.	1	.	1	1	1	.	4

Referring to information retrieval, van Rijsbergen (1979) described the completeness of the information need in traditional information systems data retrieval and then described the contemporary distinguishing properties that differentiated data retrieval from information retrieval, as seen at that time. Referring to Table 2.3, and according to van Rijsbergen (1979):

- i) The first property is *matching* – either an *exact match* or a *partial match*. In data retrieval one typically seeks for an exact match – data either exist or do not exist within a database table. In information retrieval, an exact match might be sought but often uncertainty of the user prevails and therefore a partial match becomes the actual need; thereafter the best documents returned (as judged by the user) can be selected by the user.
- ii) The second property is *inference* – reaching a conclusion based on evidence and reasoning, which may be by either deductive or inductive inference. Relationships in deductive inference can be mathematically represented as follows:

$$\text{if } a = b, \quad \text{and } b = c, \quad \text{then } a = c$$

Inductive inference in information retrieval relationships includes degrees of certainty or uncertainty and level of confidence.

- iii) The third property is the model that is applied, either deterministically or probabilistically. Data retrieval is viewed as deterministic within its processing, where all events are inevitable, while information retrieval is viewed as



probabilistic, with multiple possible outcomes, with varying degrees of certainty or uncertainty.

- iv) Classification is the fourth property and is either monothetic or polythetic. A monothetic classification suggests that all members are identical in all characteristics, for example, where classes are defined by objects that possess characteristics that belong to the class. A polythetic classification suggests that all members are similar in characteristics, but not identical, required ideally in information retrieval.
- v) The fifth property is the query language either artificial or natural in nature. An artificial query language is structured, with restricted syntax and vocabulary while a natural query language makes use of a natural linguistic language although this has its challenges.
- vi) The query specification is the sixth property and is either complete or incomplete. Often the data retrieval query specification is complete since the information need is precise, whereas the information retrieval query specification is invariably incomplete because of uncertainty.
- vii) The seventh property refers to the items required because of the information need of the user. In data retrieval efficiency is important as the user often requires the exact items (an exact match) while in information retrieval the user seeks the relevancy of an item.
- viii) Error response is the final property differentiating data retrieval from information retrieval. An error response can be either sensitive or insensitive. When errors occur in data retrieval matching, the process which is often very sensitive, aborts without returning the item, while in information retrieval the errors are more insensitive since the mismatch of the item will not affect the performance of the system significantly.

**Table 2.3: Data retrieval vs. information retrieval (Redrawn from van Rijsbergen, 1979:1)**

	<b>Data Retrieval (DR)</b>	<b>Information Retrieval (IR)</b>
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

One key property of those listed in Table 2.3 is number seven: the complex concepts of matching and relevancy. Matching refers to the exact match of a term expressed within a query to a term that exists within a document. Relevancy is based on a judgment of whether a document is relevant or non-relevant. The two types of relevancy are topical relevance and user relevance. Topical relevancy relates to when a document and a query are judged relevant with respect to a topic (subject or body of knowledge) whereas user relevancy relates to the aspects (age, language, subject matter) that assist a user to make a judgment of whether a document is relevant or non-relevant. Relevancy can also be classified as binary or multivalued. Binary relevancy relates to 0s and 1s indicating true or false, yes or no, relevant or non-relevant whereas multivalued relevancy includes additional options (or levels) to cater for the real world, for example, relevant, non-relevant, and undecided (Croft et al., 2015; Croft, 2019).

## **2.5 IRS models and methods**

Information retrieval models and methods, many of which are based on numerous mathematical formulae, provide a framework of concepts and activities that help explain information retrieval theories and assumptions. In this section, the various information retrieval theories are discussed together with numerous mathematical methods and formulae that often co-exist.

Although numerous information retrieval models are now used in research – for example, the inference network model (Tsirikika & Lalmas, 2004), probabilistic models (Robertson, 2005), the vector space model (Chew et al., 2011), the Boolean retrieval model (Croft et al., 2015), improvements to the vector space model that makes use of Precision, Recall and cosine similarity measurements (Langville & Meyer, 2007), and the term-by-document matrix (Kobayashi, Mol & Kismihók, 2015) – judging relevancy to the user (by the user) still remains a challenge. Berry et al. (1999) argue for the need for a cosine similarity threshold (a selected tolerance value) to be used when judging document relevancy. Clarke et al. (2000) suggest improvements through their short query ranking measure called *cover density* that expands coordination level ranking through the measurement of term co-occurrence. Castells et al. (2007) and Binkley and Lawrie (2015) describe the vector space model as a model whereas Langville and Meyer (2007) and Chew et al. (2011) refer to it as a method. In this study, the vector space model will be referred to as a model. The following models are now discussed: The bag of words, the citation, the Boolean, the vector space, and the Markov random model.

### **2.5.1 The bag of words model**

A bag of words refers to a collection of words that are randomly constructed in a space (called a *bag*), without structure and order. By removing a few characteristics of a language such as structure, word order, and grammar, a textual document is referred to by many authors as a bag of words and today information retrieval queries are referred to in the same way (Nguyen et al., 2018). A multiplicity of words without the characteristics of structure, word order, and grammar present a challenge when an information need arises to retrieve relevant information from within such a document. The bag of words model, referred to in the earlier work of Harris (1954), is as follows: take the ubiquitous example of the two sentences: *'Mary is quicker than John'* and *'John is quicker than Mary'* (Agnihotri, Verma & Tripathi, 2017). If the words from the sentence are dropped into a bag of words as tokens the exact ordering of the singular-word terms in a document is ignored but occurrences of each of these terms is measurable. Information is only retained on the number of occurrences for each term. Therefore, if the two sentences, *'Mary is quicker than John'* and *'John is quicker than Mary'*, are the only sentences each within their own document, then by utilising the bag of words model in information retrieval, both these documents would be treated equally (Agnihotri et al., 2017).

### **2.5.2 The citation ranking method**

To expand on the limited information that the number of occurrences of a word provides, a method of measurement called citation indexing was introduced. In the inspiring work of Garfield (1955) concerning the science of citation analysis, the foundation was laid for the ideas and concepts of the citation ranking method. These ideas were based on earlier work by Gross and Gross (1927) on citation analysis and by Pinski and Narin (1976) where the authors built on the ideas and concepts of citation analysis and ranking, and introduced a citation influence methodology for scientific publications. These ideas and concepts were ultimately commercially developed for the Web as the PageRank linking method by Brin and Page (1998), the founders of Google Incorporated. In work that is more recent, these ideas have developed into page ranked retrieval systems (Bui, Jonnalagadda & Del Fiol, 2015).

### **2.5.3 The Boolean retrieval model**

The Boolean retrieval model is one of the oldest models used by IRSs and is still in use (Yu, 2019). The Boolean retrieval model allows users to specify their information need making use of complex Boolean operators such as AND, OR and NOT, while ranked retrieval models do not (Singhal, 2001). The Boolean retrieval model uses a precise language with operators for building up query expressions, for

example,  $q_{pto} = [ Kyle AND Jane ]$  and is different from ranked retrieval models where users typically use free text queries, for example,  $q_{ftq} = [ Kyle Jane ]$ . Just before the 1990s around the advent of the Web (Berners-Lee, 1989) many IRSs made use of the Boolean retrieval model, not because of an information user's choice, but because at that time it was possibly the only feasible option. The distinct disadvantage of the Boolean retrieval model is its lack of document ranking, creating challenges in forming effective search requests (Singhal, 2001). The Boolean retrieval model makes use of inverted indices where a document either does or does not match a query; later the model was extended using additional operators, including the term proximity operator (Croft et al., 2015; Yu, 2019).

#### **2.5.4 The Vector space model**

The vector space model was first introduced by Salton et al. (1975), one of the most extensively studied models that use theories from linear algebra (Salton & Buckley, 1988; Croft, Turtle & Lewis, 1991; Van Rijsbergen, 2004; Manwar et al., 2012; Nguyen et al., 2018; Onal et al., 2018; Orkphol & Yang, 2019). It makes use of numerous theories, concepts, and measurements and is calculation intensive. It also uses inverted indices, term-by-document-matrices, tokenisation, and various frequency calculations for collections, terms, and documents together with definitive linear algebraic formulae for Precision, Recall, the F-measure, and Relevancy (Salton et al., 1975). Relevancy is referred to as the concept of similarity where a query is connected to a document. In their work, Salton et al. (1975) describe the basic workings of the vector space model and state that the vector space model is a model where queries and documents are represented by vectors and an attempt is made to match the vectors. The mathematical formulae used in the vector space model are discussed in the cosine similarity section later in this chapter, in section 2.9.8.

#### **2.5.5 Markov random field model**

From the literature, it is evident that mathematics plays a pivotal role in IRSs. A further type of mathematical model is the term dependency model. Term dependency models such as the Markov random field model, sometimes referred to as undirected graphical models, are alternatives to feature-based models that use term frequencies and document frequencies (Metzler & Croft, 2005; Chen & Welling, 2012; Hamilton, Koehler & Moitra, 2017; Brudfors, Balbastre & Ashburner, 2019; Wang, 2019). The Markov random field model has a random pattern that can be analysed statistically but maintains the effect of unpredictably. The model describes a possible sequence of events where the probability of each succeeding event depends on the state of the

preceding event and is sometimes referred to as a stochastic<sup>23</sup> method. A hidden Markov model is referred to as a statistical Markov model where the IRS is assumed to be a Markov method with hidden (unobserved) states. That said, a Markov model describes a method with a set of states with evolutions between them and for each evolution a probability exists. Figure 2.2 illustrates the four assumptions of the Markov random field model: (1) full independence; (2) sequential dependence; (3) full dependence; and (4) general dependence (Metzler & Croft, 2005; Croft et al., 2015).

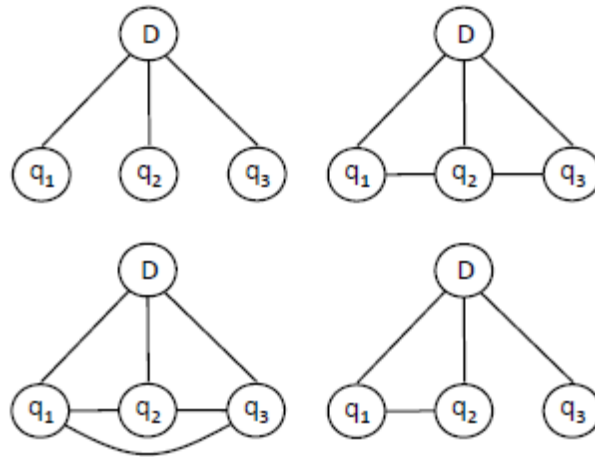


Figure 2.2: Markov Random Field model assumptions (Croft et al., 2015:455)

## 2.6 IRS Indexing methods

The concept of indexing and the need for indexing was proposed by Bush (1945) when referring to a futuristic device called a memex<sup>24</sup> that would be able to store and retrieve information from documents. An information need came about because of the need to retrieve information from huge volumes of technical, engineering and aeronautical documents soon after the end of World War II (Cleverdon, 1956) and the Uniterm indexing system, making use of a distinct list of word terms, was created to help facilitate this (Taube, 1956). The term *importance* is supported by the term *effectiveness*, representing the degree of success.

Mooers (1950; 1951) introduces the term *descriptor*, a word used to identify something, used to describe the content of a document. Mooers (1950; 1951) then coins the phrase *information retrieval*, the searching for, and retrieval of, information

<sup>23</sup> A stochastic method contains a random element that is unpredictable and exists without a stable pattern/order. It is a time sequence representing the evolution of a system represented by a variable whose change is subject to a random variation

<sup>24</sup> A memex is a conceptual data storage and retrieval system introduced by Vannevar Bush in 1945

from stored unstructured data, and introduces Zatocoding, now known as *hashing*<sup>25</sup>. Luhn (1953) theorises a new method for searching and retrieving information that provides responses in all cases not only the relevant cases, paving the way for the use of the terms *relevant* and *non-relevant*. Kent et al. (1955) introduce the concepts of *precision* and *recall* together with their mathematical formulae, and Luhn (1953) introduces the two-class classifier<sup>26</sup> known as the 2x2 contingency table (Cleverdon, 1967) or confusion matrix (Kohavi & Provost, 1998) and the F-measure mathematical formula based on precision and recall. In the four quadrant 2x2 contingency table Cleverdon (1967) makes use of *a*, *b*, *c* and *d* as measurements for each of the four outcomes placed in each quadrant, denoting *relevant retrieved*, *non-relevant retrieved*, *relevant not retrieved*, and *non-relevant not retrieved* respectively. Although Kohavi and Provost (1998) used the 2x2 contingency table, they inverted the rows and columns. However, Manning et al. (2008) use the format mirroring that used by Cleverdon (1967). This format enhances the descriptors for each of the four measurements utilising the terms: true positive (*tp*), false positive (*fp*), true negative (*tn*) and false negative (*fn*), mirroring similar concepts as in the use of the letters used by Kohavi and Provost (1998) – *d*, *c*, *b* and *a* respectively. These descriptors refer to the columns judged by the user as the truth (relevance) and the rows as the system (retrieved). The false positive concept originates from the *false drops* idea by Mooers (1950; 1951) where retrieved documents are judged non-relevant.

More recently, in-memory computing technology has become well established where databases, indices, and tables reside in memory, producing improved performance (Sadiku, Shadare & Musa, 2019). As a result of this technology, database tables can now be very large containing thousands of columns, for example, Google Bigtable. Bigtable is a distributed storage system designed to scale to a very large size using petabytes of data. The size of an index is one of the deciding factors defining the quality of Web IRSS. The index influences retrieval quality and can provide valuable Web usage information (Van den Bosch, Bogers & De Kunder, 2016; Yang, Hou & He, 2019; Sadiku et al., 2019).

The concept of text indexing dates back to Gross and Gross (1927) well before the advent of electronic computers and is based on matching words, acquired from the

---

<sup>25</sup> Each vocabulary term is hashed into an integer and at query time, each query term is hashed separately following a pointer to the corresponding postings

<sup>26</sup> A classifier is a function that takes objects and assigns them to one or more distinct classes

text termed as tokens, to a document, thus forming the basis of IRSs functionality and performance (Salton et al., 1975).

There are numerous indexing methods and a few of these are: i) the positional inverted index which stores the positions of each token that exists within a document together with the number of tokens (Procházka & Holub, 2017); ii) the inverted index where the distinct tokens are stored together with a postings list, allowing the tokens to point back to the documents they exist in (Croft et al., 2015); iii) the tiered index, also referred to as the top document index, that makes use of tiers (levels) and lists the most frequently used tokens together with their related documents at the top tier (tier-1), then tier-2, and so on; iv) the next word index that is typically a combination of indices that can identify a succeeding word (Williams et al., 2004); and v) the phrase index is the concept of multi-word indexing using two or more words referred to as *n*-grams. Although each of these indices have their own strengths there remain many weaknesses in their design: phrase-terms are not catered for in most of these, identification of each word uniquely in a collection is not possible, differentiating between both preceding and succeeding words is not possible and none of these indexing methods can accommodate phrase-term co-existence where one phrase-term co-occurs within another (Clarke et al., 2000; Manning, Nayak & Raghavan, 2017).

Salton et al. (1975) argue that indexing forms the basis of IRS functionality and performance. Four indexing theories and methods are now discussed: i) the inverted index; ii) the tiered index; iii) the phrase index; and iv) the next word index.

### **2.6.1 The inverted index**

The inverted index was the first influential concept in information retrieval, and is now the typical term used in information retrieval, but the actual term *inverted index* is obsolete as the core function of all indices is to point back from a term used in all the sections to all places where it exists within a document. The process begins by grouping the terms with naming conventions, describing this group of terms varying from a dictionary, a vocabulary, and a lexicon. Usually '*dictionary of terms*' is used for data structures (how the data are structured within the document), and a '*vocabulary of terms*' for sets of lexicon terms (a list of words used in a particular language or subject). A postings list (or inverted list) is a list of terms that exist or do not exist within a document, and if they do exist, together with their positions (i.e. the document identity of where they exist). All posting lists can then be grouped and referred to as the postings. Therefore the two components of an inverted index for a collection of documents (sorted by document identity) are: i) the postings dictionary, with the terms

typically sorted alphabetically and the postings preferably kept in computer memory; and ii) the pointers, that point to the positions in each postings list, which are often stored on a computer disc (Langville & Meyer, 2007; Mitra et al., 2017). Van Gysel (2017) provides a good explanation of how the inverted index functions. According to Van Gysel (2017), the inverted index is a specialised structure that performs term-based matching whereby a term expressed within a query is matched to a term within a document and works similarly to a subject index in a reference book.

Based within the context of the statement, Figure 2.3 is an illustration based on the theories, ideas, and concepts from Langville and Meyer (2007:3). Figure 2.3 illustrates building an index from two small documents by sorting and grouping. The sequence of terms is listed on the left-hand-side in textual sequential word order from each of the two documents,  $d_3$  and  $d_4$ , represented by their unique document identities 3 and 4 (sometimes referred to as serial numbers). From top to bottom, the first four terms relate to  $d_3$  and the remainder to  $d_4$ . Term instances in each of the two documents are alphabetically sorted and represented in the centre with their relevant document identity, for example, the term *health* exists in  $d_4$  and the term *safety* exists in both  $d_3$  and  $d_4$  and are thereafter grouped distinctly. The distinct term instances represented on the right together with their document frequencies (and possibly their term frequencies, both to be discussed in detail later in this chapter) form the term dictionary where pointers to the postings list indicate the document identities of where each term exists.

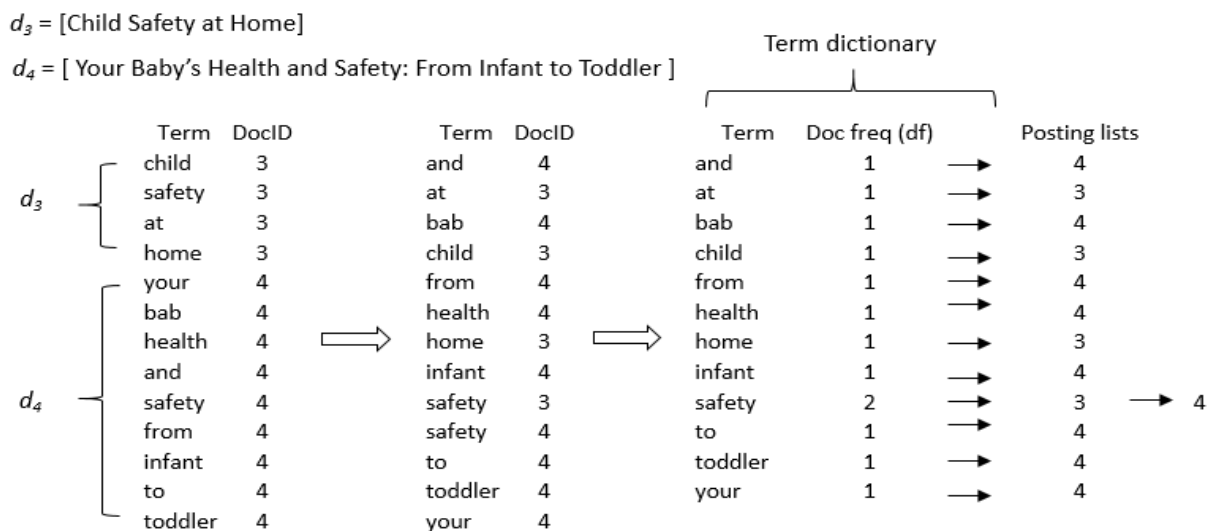


Figure 2.3: Building an index (Adapted from Langville & Meyer, 2007:3)

In their work, Alonso, Gertz and Baeza-Yates (2009) use an inverted index, where all documents are allocated a unique document identity, and the inverted index retrieves



those documents for any given query term. Egoli et al. (2000) discuss concept-based information retrieval and describe how an inverted index is used to map each word in the collection back to the concepts associated with them. Modifications to the inverted index for specific needs have been made in the past. Koopman et al. (2018) added additional fields to an inverted index to store specialised health care information. This information pertained to diagnoses, tests, and treatments.

Referring to Web-based information retrieval, Gugnani and Roul (2014) argue that the inverted index has its limitations. One limitation of inverted indices is that they are expensive to update and another is that the IRS search engine must recreate the inverted index frequently. The authors emphasise that these indices must be updated as often as possible to reflect the most recent information available on the Web thus providing the best results to the user. Van Gysel (2017) also suggests limitations with the inverted index and discusses two pointing out, that as the inverted index uses term based queries, the IRS may incorrectly judge relevant documents that do not contain query terms as non-relevant, thus affecting the recall retrieval results. In addition, search engines tend to correct indexing inaccuracies (the false negatives) by adding additional terms to the queries. In their work, Gugnani and Roul (2014) use a triple indexing method and compare it to a standard inverted indexing method, and their results show that when using an IRS search engine the query time using phrase queries is reduced by approximately fifty percent.

### **2.6.2 The tiered index**

Manning et al. (2017) provide a good explanation of how the tiered index functions. Tiered indices are typically used in Web search engines. A tiered index is essentially an inverted index that is broken up into tiers of decreasing importance. The index breaks the posting list up into a hierarchy of lists from the most important to the least important. When the search engine is fired with a query, the top tier is used. If this query fails to produce a certain term frequency threshold ( $k$ ), the index drops to its next lower level, and so on. Tiered indices are also referred to as champion lists, fancy lists, or top documents. This concept is based on setting term frequency ( $tf$ ) thresholds at various tiers (layers, levels or stratum). If the postings entries, in document order, do not match or exceed the tier (i.e.  $k$  results are unachieved) the query drops down to the next lower  $tf$  tier (Manning et al., 2017).

In their work, Rossi et al. (2013) use an indexing method that uses a two-tiered index. The first tier is a small index that contains the high impact entries, and is used to pre-process the query. The second tier is the larger index containing the details of the terms and documents. From their results, Rossi et al. (2013) posit that this two tier

indexing method speeds up the search engine and query time processes, considerably.

### 2.6.3 The phrase index

Bi-word indices refer to two single word terms used within an index. If we take the text for document  $d_5$ , “baby proofing basics”, from the example of Langville and Meyer (2007:63-3), a bi-word index could then consist of ‘baby proofing’ and/or ‘proofing basics’. The term *phrase index* is used to describe this concept of a bi-word index or a tri-word index, or multi-word indices that are greater than or equal to two words. By comparison, Shekarpour et al. (2017) refer to word terms as  $n$ -grams, single word terms as unigrams, two word terms as bigrams, three word terms as trigrams, the latter two being phrase-terms. To differentiate between single-word indices and multi-word phrase indices in this study, the descriptors *term index* and *phrase index* respectively will be used (Transier & Sanders, 2008; Wang, Huang & Feng, 2017).

Moving on from the term index and phrase index, in practice there are many complex multi-word terms referred to as compounds or phrases. Healthcare in particular uses many words and phrases emanating from Latin, and there are many technical terms in engineering, computer science and many other disciplines that make information retrieval challenging. To accommodate multi-worded terms or phrases, many search engines identify the double quotes syntax as a phrase comprising of one or more words, for example, “*Cape Peninsula University of Technology*”. Queries utilising multi-worded terms including the double quotes syntax are referred to as phrase queries and describe user acceptance of this format as successful, while multi-worded terms, without using the double quotes syntax, are referred to as implicit phrase queries. Unfortunately, postings lists of documents listing multi-word terms become inefficient, and for IRSs to be efficient they must have the capability to support these phrase queries (Transier & Sanders, 2008; Wang et al., 2017).

According to Williams et al. (2004), the phrase index is simply an inverted index where the information acquired from the text is made-up of phrases that contain more than one term rather than single-word terms. A partial phrase index only stores information pertaining to selected phrases and when the search engine is fired the IRS can only return documents efficiently pertaining to those phrases. A complete phrase index contains all phrases but in reality, it is not feasible to index all phrases owing to the index creation time and data storage requirements, among others. Williams et al. (2004) emphasise that a partial phrase index is only effective when frequently used phrase queries have been indexed. In addition, the authors suggest that, for efficiency purposes, the number of words in a phrase is recorded so that when a bi-word phrase

query is expressed, only phrases with two or more words are searched for. Williams et al. (2004) argue that a phrase containing three words cannot be used efficiently to search for a bi-word phrase.

In their research, Transier and Sanders (2008) replace the positional index with a phrase index in an in-memory IRS. The results from their experiment using two-term phrases and a two-term phrase index show significantly improved query-processing times. Transier and Sanders (2008) conclude by suggesting that if in-memory storage capacity is increased by 13% the time taken to process challenging queries could be halved.

#### **2.6.4 The positional index**

Although Transier and Sanders (2008) replace the positional index with a phrase index in their research, the positional index does have merit. Positional indexing is a more common solution than bi-word or tri-word indexing. The concept of the positional index is to store the positions of each term index, created from the tokens, that exist within a document in the format of document ID, the term frequency (number of occurrences), followed by each position (the token index) at which the term occurs within the document (Trieschnigg, 2010; Wang et al., 2017; Lahiri et al., 2019) and is presented as follows:

$$term, docid, tf: position1, position2, \dots positionn$$

For example:

$$to, 993427: (1, 6: (7, 18, 33, 72, 86, 231)$$

According to Trieschnigg (2010), the problem where multiple independent tokens may result in nondescript index terms, can be partially solved using the positional index that allows the use of phrase queries or proximity queries.

#### **2.6.5 The next word index**

Combining phrase indices and positional indices can be performed successfully. If a phrase-term becomes common through popularly used queries, then recreating the phrase index at run-time becomes inefficient. Suitable queries for combining phrase indices and positional indices are often based on recent users query behaviour. However, processing uncommon phrase queries becomes a challenge. Combination indices have been proposed in the past, for example by Williams et al. (2004), who have evaluated a method that uses indices from both phrase indices and positional indices, which they and Muller and Holzinger (2019) term the next word index.

According to Williams et al. (2004) and Kissel and Wang (2017), the next word index they have proposed uses distinct terms and posting lists and is therefore similar to an inverted index. The differentiating factors are that the next word index is a structure containing three levels tuned for the retrieval of word pairs, for example, 'to be'. The first word is denoted by  $w_1$  and the next word is denoted by  $w_2$ , where  $w_2$  is the next word of  $w_1$ .

Figure 2.4 illustrates a next word index containing two first words  $w_1$  and  $w_2$ . For each of the first word/next word pairs, a postings list is created. The first word  $w_1$  has one next word  $w_2$ . In summary for each word, the next word index stores words that follow it sequentially in a document (Kissel & Wang, 2017).

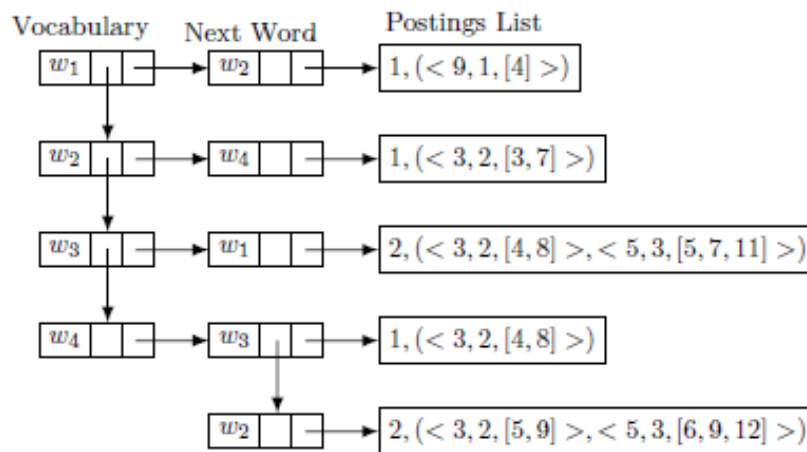


Figure 2.4: A sample next word index (Kissel & Wang, 2017:5)

In the example of a positional index, word terms and not phrase-terms are referred to; therefore, the positions that are stated refer to single words within a document. This is not a combination strategy making use of phrase-terms, as the positions would then need to cater for each word within the phrase. One combination strategy has been suggested by Williams et al. (2004) and Kissel and Wang (2017), the next word index comes close but still does not satisfy the requirements for this study.

This is a limitation in the theory of indexing and forms one of the objectives within this study. One idea would be to hybridise the features of the phrase index and the positional index and to store the unique identities of each word within a phrase-term – this idea does not appear to be in the literature. What is then needed is a hybridised index (or a combined index) that can accommodate a variable length phrase index, together with a positional index, that can store not only all the positions of the words within the phrases but can also store the sequence order of the words within each of

the phrases. This is the definitive weakness of current indexing systems used in IRSs and one of the possible root causes of the research problem.

According to Lewis (2010), a next word index uses single word tokens that are combined into word pairs. Users can be assisted in query formulation by using word pairs with high information value and high frequencies.

Janet and Reddy (2010) suggest that to overcome a few of the challenges in indexing a combination of indices should be used in the form of a cube. The model of Janet and Reddy (2010) creates a single three dimensional in cube index by using a direct index, a next word index, and an inverted index. The real world application of this indexing method can be used to cluster sets of documents based on their word proximity analysis, to browse documents rather than retrieve them, and can be used to generate rules from word associations that are contained in the index so as to enrich a user's specification.

## **2.7 IRS design concepts**

In this section, the literature is explored looking at design concepts and current theories of how to gather information during information retrieval. This is to find out who has done what and how to transform text from a document into a functioning index so that a search engine can interrogate that index to return a result. The following IRS design concepts are now discussed.

### **2.7.1 Content acquisition**

Content acquisition (Faheem, 2014) or text acquisition (Narayan et al., 2017) is the first stage of the information gathering process. Text acquisition allows available documents to be searched, effectively creating a document collection. It is basically a process of information gathering. This can be performed by scanning (crawling) the Web manually, or programmatically, to identify the documents available. The data are then passed to the index creation component through the creation of a data store (a set of text data and metadata about the document, for example, document length) (Narayan et al., 2017). The four sub-components for text acquisition are crawling<sup>27</sup>, document feeding, document converting, and document data storing. A user can manually crawl for information held within various documents or crawling can be computerised to search for document links and gather the information within the document by following the link. Information from real-time streaming can be gathered

---

<sup>27</sup> Crawling is a form of information gathering – a crawler is synonymously referred to as a spider and therefore crawling is often referred to as spidering.

through the feed mechanism. However, one critical sub-component in text acquisition is the conversion of various file formats (.docx<sup>28</sup>, .pdf<sup>29</sup>, .ppt<sup>30</sup>) into constructed text and metadata format – essentially all file formats must be converted to text format. The document data store is a database that manages large volumes of documents and the structured data associated with them. It is typically a relational database that contains the metadata from the documents collected (Faheem, 2014; Narayan et al., 2017). Information can be gathered by collecting documents containing textual information from numerous sources including: i) the Web; ii) social media; and iii) test collections. These sources are now discussed in further detail.

### 2.7.1.1 The Web

The first source of textual information illustrated is the Web. Berners-Lee (1989) is accredited with the invention of the Web at the European Organisation for Nuclear Research in Geneva in March 1989. The original motivation behind the design of the Web was to access library information via page linkages to documents, making use of hypertext<sup>31</sup> (a system that allows widespread cross referencing between related sections of text), residing on various servers (Berners-Lee & Fischetti, 1999; Berners-Lee, 2000; Berners-Lee & Fischetti, 2000). Recently, the *'Information management: a proposal'* by Berners-Lee (1989) that outlined the concepts, ideas and indexing of the Web had its 30 year celebration (Mercier, 2019). The Web thus allowed IRSs to retrieve online documents via indices.

According to van den Bosch et al. (2016), the size of an index is one of the determining factors defining the quality of Web IRSs. The index influences retrieval quality and can provide valuable Web usage information. In their work, van den Bosch et al. (2016) make use of a graph (Figure 2.5) to explain their results, from a 9-year longitudinal study of estimating Web-based information retrieval index size variability. Referring to the work of van den Bosch et al. (2016) and by comparing the indices from various IRSs, one index (Google) attained a high peak approaching 50 billion Web pages. From an estimation of 5 billion Web pages at the turn of the century (Powell, 2004), to an estimation of 50 billion Web pages in 2016 (Van den Bosch et al., 2016), the Web has become the world's largest document collection.

---

<sup>28</sup> Word – a word processing file format

<sup>29</sup> Portable document format – an electronic file format for documents

<sup>30</sup> PowerPoint – a presentation file format

<sup>31</sup> Hypertext Transfer Protocol (HTTP) can be viewed as a software application termed a Web browser that retrieves information in the form of a Web page document. A Web page is typically situated on the Web server with hyperlinks on each page allowing other Web pages to be detected. This activity is commonly known as 'surfing' or 'browsing'.

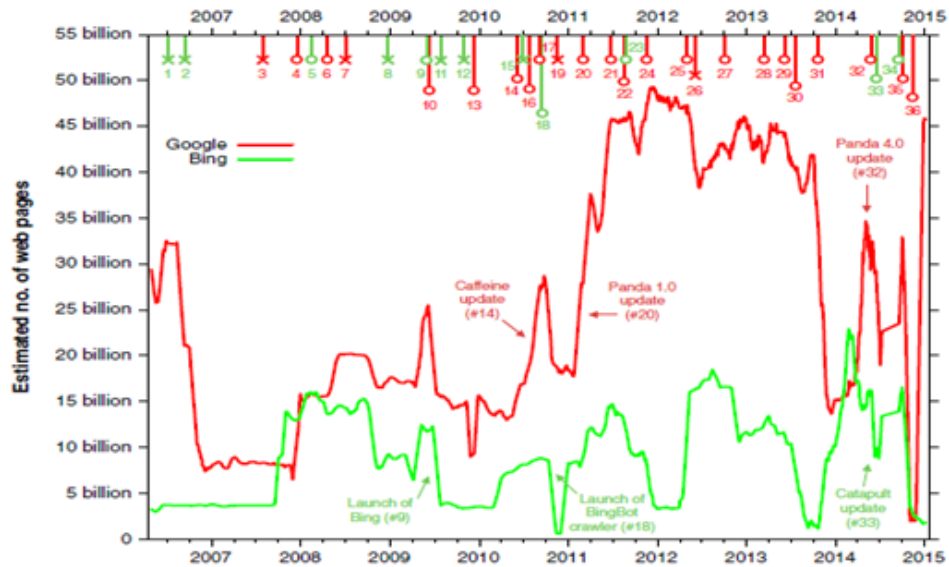


Figure 2.5: Estimated size of Google and Bing indices (Van den Bosch et al., 2016)

### 2.7.1.2 Social media

A second source of textual information is social media. Social media consists of Websites and applications enabling users not only to create content, but also to share the content thus allowing users to participate in social networking. A search application involves various communities of people that tag content or answer questions: this process is described as social search (Zamberi et al., 2018). The social media Websites that sit on the Web are referred to as *Web 2.0*, or a *Web of people*, as opposed to the traditional Web founded by Berners-Lee (1989), that consists of non-interactive documents. To differentiate between the two, the traditional Web is now referred to as *Web 1.0*. *Web 3.0* is referred to as a Web of knowledge connections and *Web 4.0* as a Web of intelligence (Patil & Surwade, 2018). There are numerous social media websites, for example, Twitter<sup>32</sup>, Facebook<sup>33</sup>, LinkedIn<sup>34</sup>, Flickr<sup>35</sup>, YouTube<sup>36</sup>, CiteULike<sup>37</sup>, Digg<sup>38</sup>, MySpace<sup>39</sup>, and others. Collectively these websites and their content provide a huge collection of textual information (Patil & Surwade, 2018; Zamberi et al., 2018).

<sup>32</sup> <https://twitter.com>

<sup>33</sup> <https://www.facebook.com>

<sup>34</sup> <https://www.linkedin.com>

<sup>35</sup> <https://www.flickr.com>

<sup>36</sup> <https://www.youtube.com>

<sup>37</sup> <http://www.citeulike.org>

<sup>38</sup> <http://digg.com>

<sup>39</sup> <https://myspace.com>

### 2.7.1.3 Test collections

Additional textual information can be sourced from test collections. Test collections that are used to experiment with using IRSs consist of three items: a collection of text documents, a sample of queries, and a list of documents judged relevant to the information need. A minimum of 50 queries that support the information needs is suggested. Test collections often change over time and reflect the adaptation to use of IRSs by users. A few of these are now discussed:

The TREC collection (abbreviated from Text REtrieval Conference) is a set of numerous test collections (figure 2.6) associated with the annual TREC evaluation forum (NIST, 2018).

## Participant Papers








-  [BJUT] BJUT at TREC 2018: Incident Streams Track  
Ning Lu, Hesong Wang, and Zhen Yang - College of Computer Science, Beijing University of Technology, China
-  [Brown] Brown University at TREC Precision Medicine 2018  
Prakrit Baruah, Riya Dulepet, Kyle Qian and Carsten Eickhoff - Brown University, USA
-  [Cat\_Garfield] Team Cat-Garfield at TREC 2018 Precision Medicine Track  
Xuesi Zhou, Xin Chen, Jian Song, Gang Zhao, and Ji Wu - Department of Electronic Engineering, Tsinghua University, Beijing, China
-  [cbnu] CBNU at TREC 2018 Incident Streams Track  
Won-Gyu Choi, Seung-Hyeon Jo, and Kyung-Soon Lee - Division of Computer Science and Engineering, CAIT, Chonbuk National University, Republic of Korea
-  [cbnu] CBNU at TREC 2018 Precision Medicine Track  
Seung-Hyeon Jo, Won-Kyu Choi, and Kyung-Soon Lee - Division of Computer Science and Engineering, CAIT, Chonbuk National University, Republic of Korea
-  [CUIS] CUIS Team at TREC 2018 CAR Track  
Xinshi Lin and Wai Lam - The Chinese University of Hong Kong
-  [DICE-UPB] DICE @ TREC-IS 2018: Combining Knowledge Graphs and Deep Learning to Identify Crisis-Relevant Tweets  
Hamada M. Zahera, Richa Jalota, and Ricardo Usbeck - Data Science Group, Paderborn University, Germany

Figure 2.6: A sample of papers from the NIST 2018 collection (NIST, 2018:1)

The TREC-AP test collection (abbreviated from Associated Press) is a collection of documents (with an average 474 words per document), supplied on three compact disks, containing 242,918 newswire documents from 1988 through to 1990 and includes the queries and relevance judgments generated by government information analysts (Croft et al., 2015).



The Cranfield-2 test and its collection is one of the original test collections for IRs that allows quantitative measures to determine effectiveness. It is a collection of 1,398 documents containing abstracts from aerodynamics journal articles, 225 queries, numerous human relevance judgments, 14,000 lines of text containing 250,000 words of which 15,000 are unique (Cleverdon, 1956; Cleverdon & Keen, 1966; Cleverdon, 1991; Scholer, Kelly & Carterette, 2010).

The CACM test collection (abbreviated from the Communications of the ACM) is a collection of bibliographic documents (Figure 2.7) containing thousands of abstracts and titles from 1958 to 2019 (ACM, 2019a).

## Communications of the ACM

### Volume 62, Number 8, August, 2019

Cherri M. Pancake	<a href="#">From the President: Dispelling common myths about ACM awards and honors</a> . . .	5--5
Vinton G. Cerf	<a href="#">Cerf's up: Undo, redo, and regrets</a> . . .	7--7
CACM Staff	<a href="#">Letters to the editor: A case against mission-critical applications of machine learning</a> . . . . .	9--9
Mark Guzdial	<a href="#">BLOG@CACM: Cutting the wait for CS advice</a> . . . . .	12--13
Samuel Greengard	<a href="#">News: The algorithm that changed quantum machine learning</a> . . . . .	15--17
Don Monroe	<a href="#">News: I don't understand my car</a> . . . . .	18--19
Gregory Mone	<a href="#">News: What makes a robot likable?</a> . . . . .	20--21
David Weintrop	<a href="#">Education: Block-based programming in computer science education</a> . . . . .	22--25
Marshall W. Van Alstyne	<a href="#">Economic and business dimensions: A response to fake news as a response to Citizens United</a> . . . . .	26--29
George V. Neville-Neil	<a href="#">Kode vicious: MUST and MUST NOT</a> . . . . .	30--31
Marco Aiello	<a href="#">Viewpoint: The success of the web: a triumph of the amateurs</a> . . . . .	32--34

Figure 2.7: An ACM collection (ACM, 2019b:1)

The GOV2 test collection (abbreviated from Government) is a collection of 25 million Web page documents from websites in the domain of '.gov' using 150 queries, during early 2004. The collection includes the queries designed as title fields and relevance judgments generated by government analysts (Hauff, 2010; Maxwell, 2014).

The NTCIR collection (abbreviated from NII Test Collections for IR Systems) is a set of test collections, in similar size to those within the TREC test collection, with the project focussing on East Asian language and cross-language information retrieval. Queries are produced in one specific language and then processed over a document collection in differing languages (NTCIR, 2019).

The CLEF test collection (abbreviated from Cross-Language Education and Function) is an evaluation series and free online resource pertaining to European languages and cross-language information retrieval (CLEF, 2016).

The Reuters-21578 test collection is the most used, known as the benchmark for text classification, and is considered a classic collection containing 21,578 newswire articles. The origins stem back to the work performed by the Carnegie Group and Reuters when developing the CONSTRUE text classification system (Korde & Mahender, 2012).

The Reuters-RCV1 test collection (abbreviated from Reuters Corpus Volume 1) is a more recent collection containing 806,791 documents, transmitted over the Reuters newswire over one year between August 1996 and August 1997, and because of its scale it is considered a much better foundation for further research than other collections (Lewis et al., 2004).

The Newsgroups test collection is an additional text classification collection. The collection consists of 1,000 articles from twenty individual Usenet newsgroups (Lang, 1995).

### **2.7.2 Text transformation**

Text transformation is the second stage of the information gathering process. In the work of Patil, Dave and Varma (2013), the authors use text transformation to transform a chunk of text into a concept space. Text transformation is therefore the component that transforms the text within documents into chunks of data, referred to as tokens. These indexed tokens are acquired from the text within the document. The set of all indexed tokens is referred to the index vocabulary. The six sub-components are tokenisation, stopping, stemming, link extraction and analysis, information extraction, and classification.

Ali (2013) refers to the process of extracting words from text as tokenisation; however, Wang et al. (2017) refer to it as parsing. In this study, the term tokenisation will be used. Short frequently used words such as *'of'*, *'and'*, *'the'* are referred to as stop words (Tijani et al., 2017), and the process is referred to as stopping (Hauff, 2010; Jimmy et al., 2018). These words are often removed from an index to save space and to increase efficiency but when these words are used in a query they will not be found. Stemming is another process that groups words derived from a common stem (Tordai, 2006). Information pertaining to the links of Web pages can be extracted and analysed

using various algorithms – this forms the basis of the PageRank method of Brin and Page (1998).

There are numerous additional methods of text transformation gathered from the literature based upon old and new theories. Seven are now discussed in further detail and these are tokenisation, Levenshtein distance, delimiters, case folding, hyphenation, suffix stripping, and stemming.

### **2.7.2.1 Tokenisation**

For efficiency and speed purposes of indexing at retrieval time, the token index must be created in advance. The first step is to collect all the documents – in this example of Langville and Meyer (2007) there are seven documents. The second step is to tokenise the text by identifying, and then acquiring, the actual words that exist within the text of the documents into chunks known as tokens (Lang, 1995). This process is sometimes referred to as chunking (Marrero et al., 2010; Adouane & Dobnik, 2017). Ali (2013) refers to tokenisation of text as the process of turning each document into a list of tokens. In the work of Stokes et al. (2009), the authors discuss tokenisation tools that can be used for this purpose and have developed their own in-house tokenisation strategy. Although the work of Ali (2013) relates to the Arabic alphabet, that has different challenges, and not the 26 letter Latin based alphabet used in English as in this research, Ali (2013) states that the simplest way to tokenise text is to use the white spaces preceding and succeeding the chunk of text. However, Marrero et al. (2010) argue that one cannot rely on these white spaces or special characters as delimiters for chunking. In their work, Marrero et al. (2010) discuss the issue of white spaces, that ‘gap’ between each token, and Ali (2013) suggests the removal of punctuation (or special) characters, for example, ‘ “ / - @ \$ : % ^ & . Ruthven and Lalmas (2003) and Agnihotri et al. (2017) suggest that all tokens are converted to lowercase and that these special characters are removed during the tokenisation process.

### **2.7.2.2 Delimiters**

Delimiters are characters that identify the beginning and/or the end of a chunk of text. Delimiters have been used in Internet email messages indicating the beginning and the end of the text message (Partridge, 2008) and can be rule-based when looking for patterns in Hypertext Markup Language (HTML) pages (Chang et al., 2006). In information retrieval working with text documents there is a need to identify the beginning and end of a chunk of text during the tokenisation process. Marrero et al. (2010) posit that white spaces cannot be relied on as delimiters of text and some form of special character should be used. Baxter et al. (2007), Farwick et al. (2013) and

Ayumba (2015) refer to files that contain text and delimiters as comma-delimited text files that use the comma character ‘,’ as a delimiter, and these are known as Comma Separated Value (CSV) files. A good example of the use of CSV files in health care is the work of Troshin et al. (2011) where for their protein information management system the authors use CSV files for uploading and downloading data for Deoxyribonucleic acid (DNA) sequencing. However, in his work entitled, ‘*The structure of science information*’, Harris (2002) discusses the representation of text and sentences and suggests the pipe<sup>40</sup> ‘|’ character (or vertical bar as it is sometimes referred to), rather than the comma character, be used to replace these whites paces, as illustrated in Figure 2.8.

Formulaic representation of sentences

It seems clear from all the evidence that the cells responsible for the synthesis of antibody shortly after the injection of a second antigenic stimulus are members of a family which arise from some undifferentiated precursor as the direct result of the stimulus.	It seems clear from all the evidence that the cells   are   members of a family WH     antigen   the injection of the second stimulus of    shortly after    antibody   (are) responsible for the synthesis of   (cells) ← which     the stimulus    as the direct result of    (Members of a family)   arise from   some undifferentiated precursor	M C <sup>W</sup> YC <sup>hw</sup> GJ <sup>2</sup> : <sup>e</sup> A V <sup>p</sup> C GJ <sup>2</sup> :C <sup>l</sup> Y <sup>t</sup> C <sub>b</sub>
The first cells which demonstrably contained antibody and can therefore be assigned to this family are large cells with a thin rim of basophilic cytoplasm and large nuclei whose appearance is indistinguishable from that of other primitive hematogenous cells.	The cells   are   large cells Which     (antigenic stimulus)   (the second injection of)    first (after)   antibody   demonstrably contain   (cells) ← and therefore (which)     (cells)   can be assigned to   thjs family WH     (large cells) with a thin rim of cytoplasm (which)   (is) basophilic WH     (large cells with) nuclei (which)   (are) large whose     (large cells’)   appearance is indistinguishable from that of   other primitive hematogenous cells	C <sup>W</sup> YC <sup>hw</sup> GJ <sup>2</sup> : <sup>e</sup> AV <sup>i</sup> C CYC <sup>l</sup> C <sup>g</sup> S <sub>c</sub> <sup>-</sup> W <sub>g</sub> C <sup>e</sup> S <sub>n</sub> W <sub>g</sub> C <sup>e</sup> YC <sub>b</sub>
During the 2 or 3 days after their first appearance they multiply, synthesize antibodies specific for the antigen which stimulated their development, and differentiate through immature to mature plasma cells.	the large cells   multiply,     (antigen)   (was twice injected)    WH ←    antibody specific for the antigen   synthesize   (the large cells) ← which     (antigen)    stimulated    the large cells’   development, and     (the large cells’)   differentiate   through immature (plasma cells)   to mature plasma cells during the 2 or 3 days after     (antigenic stimulus)   (a second injection of)    (at a time which was) first (after)    the large cells’   appearance	C <sup>g</sup> W <sub>p</sub> G <sup>W</sup> J <sup>2</sup> :A <sup>G</sup> V <sub>p</sub> C <sup>g</sup> G:C <sup>g</sup> W <sub>p</sub> C <sup>g</sup> Y <sub>c</sub> <sup>ft</sup> C <sub>z</sub> <sup>m</sup> C <sub>z</sub> <sup>m</sup> GJ <sup>2</sup> : <sup>e</sup> C <sup>g</sup> W <sub>i</sub>

Figure 2.8: Representation of sentences (Harris, 2002:5)

### 2.7.2.3 Levenshtein distance

In the work of Levenshtein (1965) entitled, ‘*Binary codes capable of correcting deletions, insertions, and reversals*’, the author introduces a measurement between two strings now known as the Levenshtein distance. The Levenshtein distance or edit distance described by Gusfield (1997) is a similarity measurement made between two character strings, say  $s_1$  and  $s_2$  (Manning et al., 2008). The Levenshtein distance is therefore defined as the least number of edit operations (deletions, insertions, or replacements) required to transform a character string  $s_1$  into  $s_2$ . For example, the

<sup>40</sup> The pipe delimiter is the preferred delimiter in information systems data retrieval processes where data are extracted from tables of a legacy information system and converted to files that contain text. A delimiter is used to separate the data in textual format emanating from the table columns; Microsoft traditionally use a comma as a delimiter in their comma separated values file (csv) formats but a comma often exists within data causing data misalignment in the textual output.

Levenshtein distance to transform 'Kyle' into 'Jane' is three. The three characters 'k', 'y', and 'l', all need to be replaced at least once with 'j', 'a' and 'n'. Weights can be applied to each of the edit operations under certain circumstances but in general edit operations are weighted equally (Levenshtein, 1965; Gusfield, 1997). Kadous (2002) argues that the Levenshtein distance between two strings is the minimum number of differences between them, and in more recent work in automated mapping of clinical terms in the health care domain, Allones, Martinez and Taboada (2014) posit the Levenshtein distance as a type of edit distance that measures similarities between two strings of text.

#### **2.7.2.4 Case folding**

According to Ruthven and Lalmas (2003) and Agnihotri et al. (2017), during tokenisation, a typically utilised strategy is reducing all letters in the text to lowercase, known as case folding. This will allow query terms to match textual terms, for example, matching 'Idiopathic', with an uppercase 'I' at the beginning of a sentence, with 'idiopathic' with a lowercase 'i' contained within a query. Bell et al. (1993) warn that there are disadvantages when using case folding where vocabularies change. For example, it is sometimes better to keep uppercase context when attempting to differentiate between say company names and other words spelled identically, for example, 'Delta Motors' and the words 'delta' and 'motors'. By removing punctuation and special characters during tokenisation, acronym normalisation occurs, for example, 'C.A.T.' reduced to 'CAT', and if lowercased then case folded to 'cat'. Truecasing determines correct word capitalisation when information is unavailable, for example, capitalising the first word in a sentence. Truecasing can be used as an alternative to lowercasing where machine learning decision techniques are applied, but in practice, lower casing is often the most suitable. Procházka and Holub (2017) concur with Ruthven and Lalmas (2003) about the suitability of lower casing and use case folding as a process to populate their positional inverted index.

#### **2.7.2.5 Hyphenation**

According to McCray (1998), Markey (2009), and Waitelonis (2018), hyphenation – the use of the hyphen ('-') – a punctuation mark used to join words, complicates information retrieval because of its three main purposes within the English language. These are: i) splitting up vowels in words (e.g. 'coexisting' versus 'co-existing'); ii) joining nouns as names (e.g. 'Mercedes Benz' versus 'Mercedes-Benz'); and iii) copyediting (the process of improving text formatting, style, and accuracy) to illustrate word grouping (e.g. *term by document* versus *term-by-document*). Reviewing the three examples above, one could argue the first should be regarded as a single token,

the second as indistinct, and the last to be kept as separate words. Markey (2009) explains the use of hyphenations, and that hyphenation has different meanings in different languages and that each language has its own set of hyphenation patterns. McCray (1998) posits that the use of hyphenation in text changes the way multi-word phrases are expressed. McCray (1998) uses the example of 'fire power' and explains it can be written as 'fire-power' with the hyphen and 'firepower' as one word. These differentiations in writing phrases create challenges when attempting to tokenise the text.

#### **2.7.2.6 Suffix stripping**

According to van Rijsbergen (1979), Adouane and Dobnik (2017), and Waitelonis (2018), suffix stripping is a process of removing suffixes from a word to obtain the stem and warns that this is a complicated process. Because of this complicated process, a standard approach is often used whereby a complete list of suffixes is produced and the longest one is removed. However, van Rijsbergen (1979) warns that context free removal can create substantial errors, since suffixes not intended to be removed, are often removed. Often context rules are created and then applied to ensure the rule is only applied when the context is right. This 'right' has two meanings according to van Rijsbergen (1979): the first is that the length of a remaining stem may exceed a given number – a default of two is usually used; and the second, the end of the stem may satisfy a specific condition, for example, the last character is not a 'q'. However, there are methods available to strip these suffixes, as the one presented by Porter (1980). Porter (1980) emphasises that as a document can be represented by a vector of words, known as terms, and when the original position of the term is ignored (he provides the example of five terms: connect, connected, connecting, connection, connections), the five terms all have similar meaning through their common stem or root, *connect*. Reducing these five terms to a single term, *connect*, by removing the various suffixes of: -ed, -ing, -ion, and -ions, Porter (1980) suggests that an IRSs performance can be improved, and since the number of terms have been reduced from five to one, database size and complexity can be reduced.

#### **2.7.2.7 Stemming**

Tordai (2006), Halácsy and Trón (2007), and Frej, Chevallet and Schwab (2018) suggest that stemming could be a solution to improving information retrieval. Morphological tools encompass both stemming algorithms and lemmatises. Stemming allows words to be reduced to their root (or stem) through the removal of affixes – prefixes and suffixes (Tordai, 2006). The benefits of stemming include reduced time in, increased volumes of, and precision of, information retrieval (Tordai,

2006). Search engines that treat words with the same stem, as the same, follow a process known as conflation (Porter, 1980). Conflation occurs when two items with similar characteristics are treated equally (or morphed, as if they were the same).

In the works of Palangi et al. (2016) and Frej et al. (2018), neither set of authors used any form of stemming on their document collections as they concur with the opinion that the best results are obtained without stemming. In addition, Palangi et al. (2016) retained numbers, used white spaces as delimiters for tokenisation and case folded the text.

### **2.7.3 The data store**

Pinkerton (2000), Troshin et al. (2011), and Voorslys, Broberg and Buyya (2011) have all made use of data stores for their research. Pfeiffer et al. (2008) discuss the role of the data store in Genome information management. A Web application is the primary interface and is connected directly to the data store. This allows browsing and the querying of data with minimal effort over the Web. In addition, access rights are managed, thus allowing the management of data within the same data store. This process, of hosting the token index within the data store, by design, should be efficient in time, space, and updating.

### **2.7.4 The token index**

Token index creation is the third stage of the information gathering process. The token index takes the output from the text transformation stage and creates the index in the data store with these tokens. The token index will typically contain all the tokens and the document numbers where each token exists (Liao et al., 2019).

## **2.8 Search engines and queries**

In this section, the literature is explored in order to look at current theories and methods to understand more clearly what has been done in the design of the search engine process and what makes the search engine retrieve those documents being sought. This section therefore explores methods for processing a user's query. These queries contain single-word or multi-word terms, which are presented to the IRS in an attempt to match those terms expressed in a query to the tokens within the text of the document.

### **2.8.1 Query expansion**

Query expansion is a method that uses multiple terms expressed in a single query in an attempt to improve document retrieval performance (Zhao, 2012; Scells, Zuccon & Koopman, 2019). If query results are of poor quality, queries can be revised using

query expansion by adding additional terms to the query. These expanded queries can be used to overcome the challenge of synonymic<sup>41</sup> and homonymic<sup>42</sup> lexemes<sup>43</sup> and other words within the English language that exist in many documents and it is one way of attempting to solve the vocabulary mismatch problem (Koopman et al., 2018). In query expansion, the query is reformulated by the user by using additional single-word terms or multi-word phrase-terms, in an attempt to increase the quality of the query, thus increasing the effectiveness of the IRS (Tolias & Jégou, 2013). However, expansion does not always provide an increase in IRS effectiveness, according to He and Ounis (2009) who investigated the ineffectiveness of query expansion. He and Ounis (2009) argue that this ineffectiveness is based on IRSs that retrieve too many non-relevant documents, as too many expanded terms are expressed within the query. In addition, although the IRSs retrieve these documents, they are often irrelevant to the user's information need (refer Section 2.4) and He and Ounis (2009) conclude that this makes the use of expanded terms and the use of query expansion problematic. However, in the work of Hanbury et al. (2014), the authors reviewed seven papers pertaining to intellectual property within the legal profession. Three of the seven papers, or forty-three percent, suggested using query expansion while one paper considered the use of multiple query expressions to improve patent search retrieval results. Pal et al. (2015) investigated methods using query expansion that automatically classified a given query into one or more pre-defined categories. In the conclusion of their work, Pal et al. (2015) propose a taxonomy of query classes and recommend that from a query expansion perspective, query categorisation should be a consideration. Pal et al. (2015) and Koopman et al. (2018) concur with the view that even with query expansion the problem of mismatching vocabulary still remains a problem.

### **2.8.2 User relevance feedback**

According to Hamid (2017), relevance in IRSs defines how well retrieved information meets the requirements of the user. Queries can be fine-tuned using a form of relevance feedback. The idea is to use information provided by the user on how he or she judged documents as relevant and to use this information to fine-tune the queries in an attempt to improve the results. Automating the manual part of relevance feedback is possible through pseudo relevance feedback, sometimes referred to as

---

<sup>41</sup> A word having nearly the same meaning as another word

<sup>42</sup> Two or more words that have the same spelling but different meanings

<sup>43</sup> A basic unit of a language consisting of one or more words



blind relevance feedback. It is an automatic local analysis that improves information retrieval performance (Hamid, 2017; Wang et al., 2017).

Rocchio (1965) argues that relevance feedback in IRSs is an iterative process whereby a user can fine tune queries in an attempt to retrieve those documents that are relevant. The process begins, firstly, with a query that is communicated to the IRS and a set of documents is then retrieved by the system. Secondly, the user can then intervene and judge whether the retrieved documents are relevant or non-relevant. Thirdly, based on the feedback of the user considering their information need, the system then typically provides an improved representation, through a revised set of retrieved documents. There are numerous methods to measure relevance feedback, for example the Rocchio algorithm (Rocchio, 1965; Rocchio & Salton, 1965; Hamid, 2017) that includes relevance feedback information into the vector space model.

### **2.8.3 Ranked retrieval**

Ranked retrieval is the process of ranking IRS retrieved results based on a specific parameter. Ranked retrieval applies to an ordered set of documents. It is dissimilar to an unordered set of documents where set-based measurements are used, for example, in Precision, Recall, and the F-measure. The basic concept of ranked retrieval is for an IRS to return the top- $k$  documents, where these documents should be the most appropriate (Kelly, 2009; Trieschnigg, 2010; Mao et al., 2015).

In ranked retrieval, the precision-recall curve illustrates the precision and recall values and it is often zigzag shaped or in the shape of a saw-tooth. The underlying principle for this curve is relevance and when the  $k + 1$  document is retrieved and is judged non-relevant, then the value for Recall for the top- $k$  documents is equal. But as the  $k + 1$  document retrieved is non-relevant, this affects precision and therefore precision decreases. However, if the  $k + 1$  document retrieved is relevant then both Recall and Precision increase creating the zigzag shape (Kelly, 2009; Mao et al., 2015; Sankhavara, 2018). Nevertheless, in some document collections, fewer than the top- $k$  documents are retrieved as in the example by Langville and Meyer (2007).

Based on citation analysis, the PageRank algorithm was introduced by Brin and Page (1992). The basic workings of this algorithm, used by Google Incorporation, is based on the Web page link structure. If a Web page is judged important, and it has forward-links to other Web pages, then these additional Web pages are judged important. The return-links, of these Web pages, are used by the PageRank algorithm to determine the ranking and provides a score.

The Hyperlink-Induced Topic Search (HITS) algorithm is an iterative algorithm presented by Kleinberg (1999). This algorithm analyses Web page links and then ranks them by identifying two different Web page forms: Hub and Authority. Hub Web pages are resource lists and supply information to users about authoritative Web pages. This algorithm is based on the relationship between relevant authoritative Web pages and hub Web pages. These Web pages are then joined together in a link structure.

#### **2.8.4 Word proximity and ordinality**

According to Weideman (2001), ranking refers to how an IRS provides the results from a user's query. Referring to ranking and proximity searching, Keen (1992a; 1992b) creates a ranking measure by weighting the distance between proximate terms. Therefore, a form of ranking is term proximity (Mitra, Diaz & Craswell, 2017), a form of distance measurement between terms in a document. When query terms occur close together, an IRS that uses term proximity allocates a higher score to that document. Therefore, one way of increasing IRS search efficiency is through term proximity where the individual terms within the multi-worded term document text appear close together. Earlier, in section 2.5.3, Boolean retrieval models were discussed and how they incorporated additional operators including the term proximity operator or closeness. Closeness and term proximity operator specify that two terms within the query must exist close to each other within the document (Skrlj, Martinc & Pollak, 2019). Term proximity weighting applies to multi-worded phrase queries where the terms within the query occur close to each other within the document text. Term proximity applies to the distance between query terms (the query must have at least two terms) that exist within a document. Taking document  $d_4$  from the Langville and Meyer (2007) example where  $d_4 = \text{'Your Baby's Health and Safety: From Infant to Toddler'}$  and using the query  $q_2 = \text{'infant toddler'}$ , the smallest window ( $\omega$ ) within the document is three and therefore as  $\omega$  decreases, the query  $q_2$  to document  $d_4$  match increases. Ideally, this is used when all query terms exist within a document, but in the situation when not all query terms exist within a document,  $\omega$  must be set to a large number. The concept of term proximity weighting can be looked at with the first term *'infant'* in position 7 of the text and the second term *'toddler'* in position 9 of the text. To calculate the window  $\omega$ , the number of words between the begin and end positions must be calculated. In this example, all query terms exist within the document. For  $q_2$ :  $\omega = 3$ , and the value of  $\omega$  is very low, while for  $q_3 = \text{'your toddler'}$ :  $\omega = 9$  and  $\omega$  is higher. Ideally, for a two-term query,  $\omega$  should be 2 (without any words occurring in-between) (Langville & Meyer, 2007).

Term proximity and word proximity are synonymic phrases that are interchangeable (Maxwell, 2014). When a query, containing two or more terms or words is expressed in an IRS, documents are retrieved by the system that contain these words but word ordinality and word proximity are lost (Clarke et al., 2000). Word ordinality can be lost through stopping and stemming (Porter, 1980) and the proximity between these words can be vast as these words can appear separately on different pages within a document, thus reducing efficiency.

One way of increasing efficiency is through the proximity search. In their work, Brin and Page (1998) introduce the Google Web search engine and emphasise the use of proximity search features where the system locates information of all Web page hits and makes extensive use of proximity search as the proximity information helps increase relevance for search queries. The authors argue the importance of the proximity of word occurrences and discuss multi-word term searches and the complications encountered when using proximity search. Brin and Page (1998) state that those hits occurring close together within a document are then weighted higher than those hits occurring further apart. Proximity is computed, based on how far apart the hits are in the document, for every matched set of hits. According to Gupta (2008), one theoretical method of proximity search is the  $k$ -word proximity search. It is used to determine whether two words are adjacent to each other and defines the  $/k$  operator as  $word1 /k word2$  where  $k$  is a positive integer argument. The  $/k$  operator is used to determine the occurrences of  $word1$  within  $k$  words of  $word2$  so if it is required that  $word1$  is to be adjacent to  $word2$ , or if  $word1$  is in position  $p$  then  $word2$  must be in position  $p + 1$  or  $p - 1$ , then  $k = 1$ . Google uses a function called `AROUND()` that takes a number, say  $k$ , indicating how many words can separate two sets of terms in a user's query (Chitu, 2010) and this type of proximity search indexing, according to Gupta (2008), is most useful for the user, as all the term combinations in documents allow the user to retrieve documents quicker.

According to Wilkinson, Zobel and Sacks-Davis (1995) and Rose and Stevens (1996), when short queries are processed, the user's expectation is to receive documents ranked firstly by documents that contain all the query terms, secondly by documents containing most of the query terms, and lastly, documents that contain only a few query terms, irrespective of term frequency (the number of occurrences of a term that exists within a document). Wilkinson et al. (1995) argue that this expectation by the user for short queries, containing between two and ten terms, remains in force even when a document with fewer terms is judged relevant, while a judged non-relevant

document contains more terms. Rose and Stevens (1996) state that of primary importance is the number of matches of query terms to documents.

For ranking short query results, the major factor is the number of matching query terms that exist within a document; this is known as the coordination level (Wilkinson et al., 1995; Rose & Stevens, 1996). Clarke et al. (2000) argue that modifications to coordination level ranking could be more effective than the relatively poor performance of the traditional cosine similarity measure used in the vector space model. Therefore, in their work the authors introduce a short query ranking measure called *cover density* that expands on coordination level ranking through the measurement of term co-occurrence, and rank documents within a coordination level based on the proximity and density of query terms existing within documents. Each document within the collection is handled as an ordered sequence of terms, the sequential order in which a user would read the words within a document.

Clarke et al. (2000) introduce the concept of a *cover* for a term set,  $T$ , for an ordered pair  $(p, q)$  referred to as the extent  $(p, q)$ , where  $t_p$  and  $t_q$  specify the start and end term positions with the interval of text beginning at  $t_p$  and continuing through to  $t_q$ , and the cover set, a collection of covers (ordered pairs) for a multi-worded phrase query. The authors make two assumptions in their work and these are: i) the shorter the cover (i.e.  $t_q - t_p + 1$ ) the higher the probability the resultant text is relevant; and ii) the higher the number of covers determined for a document, the higher the probability the document is relevant. There are a few rules too: i) no more than one cover can have the same start position (as one cover would exist within another) and covers are ordered by their start positions; ii) no more than one cover can have the same end position and are ordered by their end positions; iii) cover density is measured by the frequency and proximity of the co-occurrence of the individual query terms (Clarke et al., 2000).

Pausing for a moment, there is a need to clarify the use of various information retrieval terms within the literature. A single word is defined as a term, and terms are used within an index. In information retrieval, the goal is to match these single-word terms (word terms) with single words known as tokens within a document text. Using a poem called '*Erosion*' by Pratt (1931:1) (Figure 2.9), reference is made to one, two or three term queries and an example the authors provide, called a term set, is:  $T' = \{ "sea", "thousand", "years" \}$ . This example of a term set is clearly a set of three single-word terms and is not a set of multi-word phrase-terms, for example,  $T' = \{ "sea side", "one thousand acres", "forty nine years" \}$ .

**Erosion<sup>1</sup>**

It<sup>2</sup> took<sup>3</sup> the<sup>4</sup> sea<sup>5</sup> a<sup>6</sup> thousand<sup>7</sup> years,<sup>8</sup>  
 A<sup>9</sup> thousand<sup>10</sup> years<sup>11</sup> to<sup>12</sup> trace<sup>13</sup>  
 The<sup>14</sup> granite<sup>15</sup> features<sup>16</sup> of<sup>17</sup> this<sup>18</sup> cliff,<sup>19</sup>  
 In<sup>20</sup> crag<sup>21</sup> and<sup>22</sup> scarp<sup>23</sup> and<sup>24</sup> base.<sup>25</sup>

It<sup>26</sup> took<sup>27</sup> the<sup>28</sup> sea<sup>29</sup> an<sup>30</sup> hour<sup>31</sup> one<sup>32</sup> night,<sup>33</sup>  
 An<sup>34</sup> hour<sup>35</sup> of<sup>36</sup> storm<sup>37</sup> to<sup>38</sup> place<sup>39</sup>  
 The<sup>40</sup> sculpture<sup>41</sup> of<sup>42</sup> these<sup>43</sup> granite<sup>44</sup> seams,<sup>45</sup>  
 Upon<sup>46</sup> a<sup>47</sup> woman<sup>48,s</sup> face.<sup>50</sup>

—E.<sup>51</sup> J.<sup>52</sup> Pratt<sup>53</sup> (1882<sup>54</sup>–1964)<sup>55</sup>

Figure 2.9: A document text example (Clarke et al., 2000:4)

In this example and using the term set

$$T' = \{ "sea", "thousand", "years" \}$$

This provides a sample of a cover set as:

$$\mathcal{C}' = \{ (5, 8), (10, 29) \}$$

where the first cover is (5, 8) and the second is (10, 29); 5 is the position of the token 'sea' within the text, 8 as the position of the token 'years' within the text, and similarly, 10 is the position of the token 'thousand' and 29 as the second occurrence and position of the token 'sea' within the text. The first cover (5, 8) thus represents the text of 'sea a thousand years', as this is reduced to the four tokens of 'sea', 'a', 'thousand' and 'years' which are in positions 5, 6, 7 and 8.

Similarly, the second cover (10, 29) represents the text of 'thousand years to trace The granite features of this cliff, In crag and scarp and base. It took the sea', as this is reduced to the 20 tokens, which are in positions 10, 11 ... 29. The cover length is then computed as  $t_q - t_p + 1$  the start and end positions, and therefore the cover length for  $cover(5, 8) = 8 - 5 + 1 = 4$  and for  $cover(10, 29) = 29 - 10 + 1 = 20$ .

## 2.9 IRS matching and processing

To evaluate an IRS, the three most widely used performance measurements are Precision, Recall, and F-measure. These are best explained using a 2x2 contingency table (section 2.10, Table 2.5 and Table 2.6).

### 2.9.1 The term-by-document matrix

The term-by-document matrix is introduced as an  $m \times n$  matrix, where the rows are  $m$  and the columns are  $n$ . The term-by-document matrix is used to represent a document collection  $A$ , where traditionally the  $m$  rows represent the terms and the  $n$  columns represent the documents that are utilised to index the document collection. Each of the elements within the matrix provides a measure of importance of term  $t$  with respect to document  $d$  within the document collection (Salton et al., 1975; Langville & Meyer, 2007).

In the real world, the terms and documents represented in the  $m$  rows and  $n$  columns respectively are sometimes reversed to form an  $n \times m$  matrix by various authors for technological and practical reasons and the matrix is then referred to as the document-by-term matrix (Kobayashi et al., 2015). In large collections, the number of columns is often greater than the restricted logical limits<sup>44</sup> for various databases. Therefore, a few authors use the term '*document-by-term*' matrix synonymously with the *term-by-document* matrix (Langville & Meyer, 2007; Kobayashi et al., 2015).

According to Langville and Meyer (2007), most IRSs that make use of traditional document collections use a form of the vector space model developed by Gerard Salton (Salton et al., 1975; Salton & Buckley, 1983). In the vector space model, textual data are transformed into numeric vectors and matrices, so that key features can be discovered utilising matrix analysis techniques. In the work of Langville and Meyer (2007), the authors provide a brief example with limited explanations of the many steps that need to be followed for the vector space model.

In this section, the example is reworked and it includes comments for each of the steps. According to Langville and Meyer (2007) for a given document collection  $A$ , utilising a dictionary of  $m$  terms, document  $i$  is represented by an  $m \times 1$  document vector  $d_i$ . Document collection  $A$  is then represented by a single row, of which the columns are the document vectors.

$$A = [d_1 \ d_2 \ \dots \ d_n]$$

### 2.9.2 Term frequency

This is not a binary representation with the numerical values of 0 and 1. It is a coincidence that no more than one term exists in a document within this example from

---

<sup>44</sup> The Microsoft Access database has a logical database limit of 255 columns, Oracle has a limit of 1000 and SQL-Server 1024 but Cloud Bigtable can accommodate thousands of columns.

Langville and Meyer (2007). The term-by-document matrix (Figure 2.10) can be populated with values greater than or equal to 1, referred to as the term frequencies ( $tf$ ) used in the vector space model. These theories are echoed by Singhal (2001), where he argues the significance of term weight formulation where words often occur within a document. Term frequency ( $tf_{t,d}$ ) of term  $t$  in document  $d$  is defined as the number of times that term  $t$  occurs in document  $d$ . In the example  $tf_{t1,d2} = 1$  and  $tf_{t4,d3} = 0$  with subscripts denoting the term and document order respectively (Agnihotri et al., 2017).

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Figure 2.10: A term-by-document matrix (Langville & Meyer, 2007:63-3)

In the Boolean model, the model records whether a term is present or absent. However, in information retrieval a weighting scheme for the term frequency is required to collect further evidence and to provide more weight to a document that contains the term within its text numerous times, than to those documents where the term exists only once (Manning et al., 2008). These weighting schemes, when building the term-by-document matrix, are used in practice rather than raw frequency counts (collection frequency) to improve performance. Figure 2.10 illustrates the  $9 \times 7$  term-by-document matrix for document collection A for this example of Langville and Meyer (2007).

### 2.9.3 Document frequency

The document frequency ( $df_t$ ) is defined as the number of documents in which term  $t$  occurs and discloses information through the inverse measure of the informativeness of term  $t$  (Rennie & Jaakkola, 2005; Agnihotri et al., 2017).

### 2.9.4 Collection frequency

The collection frequency ( $cf_t$ ) of token  $t$  is defined as the total number of times term  $t$  appears in the document collection A. Both the document frequency and collection frequency are related to each term  $t$  (Hiemstra, 2000).

### 2.9.5 Documents in a collection

The total number of documents  $N$  is defined as the total number of documents in collection A, where in this example  $N = 7$  (Hiemstra, 2000).

### 2.9.6 Inverse document frequency

When considering term weight formulation words appearing in many documents are considered common, making document content incomparable. The inverse document frequency ( $idf_t$ ) was then proposed by Karen Spärck Jones in the early 1970s (Spärck Jones, 1972; Hiemstra, 2000, Singhal, 2001; Agnihotri et al., 2017) and is a direct measure of the informativeness of term  $t$ . The  $idf$  weight of term  $t$ ,  $Log \frac{N}{df_t}$  replaces  $\frac{N}{df_t}$  to numb the effect of  $idf$  and is defined as  $idf_t = Log \frac{N}{df_t}$ , where, in this example,  $N$  is the number of documents in collection A (Manning et al., 2008).

### 2.9.7 Document weight

Assumptions for  $tf_{t,d} \times idf_t$  are grounded on two empirical observations, according to Lang (1995), with regard to the text within a document: the more occurrences a word term  $t$  has in a document  $d$ , referred to as  $tf_{t,d}$  the greater the probability the word term  $t$  is relevant to the subject contained with  $d$ ; and the more occurrences of word term  $t$  there are within all documents  $df_t$  the lower the quality of the word term  $t$  to differentiate between the documents. To determine the weight for a given document, the two features  $tf_{t,d}$  and  $df_t$  are shared by multiplying  $tf_{t,d}$  by the inverse of  $df_t$  for each word term as  $tf_{t,d} \times idf_t$ . This modification was introduced by Spärck Jones (1972). Thereafter the logarithms for  $tf_{t,d}$  and  $df_t$  are used to numb the effect of weight for larger values. In the Newsgroups test collection experiment, Lang (1995) used the following formula (Equation 2.1) to weight word term  $t$  to document  $d$ .

$$w_{t,d} = tf_{t,d} \times Log \frac{N}{df_t}, \text{ where } N = \text{no of documents in collection}$$

Equation 2.1: Weight

The document weight,  $w_{t,d}$  is assigned for each term  $t$  in each document  $d$ , and is defined as:  $w_{t,d} = tf_{t,d} \times idf_t = tf_{t,d} \times Log \frac{N}{df_t}$  and therefore  $w_{t,d} = tf_{t,d} \times Log \frac{N}{df_t}$  as the inverse document frequency  $idf_t = Log \frac{N}{df_t}$ . The objective of the document weight ( $w_{t,d}$ ) is to increase the occurrences within a document for the term frequency ( $tf_{t,d}$ ) component, and for the inverse document frequency component ( $idf_t$ ) to increase the rarity of term  $t$  in the collection. The whole idea behind document weighting is to



assign a term  $t$  within a document  $d$  so that: i) the weight is low if the term appears in most of the documents; ii) it controls the relevance signal when the weight is lower for two conditions – when  $t$  occurs in few documents or  $t$  occurs in many documents; and iii) when  $t$  occurs often in a few documents the weight is highest, thus increasing the discriminatory relevance power of them (Hiemstra, 2000, 2009).

### 2.9.8 Cosine similarity theory

Langville and Meyer (2007) argue that a document collection, a set of documents, can be regarded as a set of vectors in a vector space. Within this vector space lie two axes. When using vectors the sequential ordering of the terms in each document is lost, in the same way as in the bag of words model, together with the ubiquitous example of the two sentences: '*Mary is quicker than John*' and '*John is quicker than Mary*' (Harris 1954; Agnihotri et al., 2017). Measuring similarity between two documents in a vector space is challenging because of the effect of document length. For example, if we take two documents with dissimilar document lengths (one is much longer than the other) but similar in content, the vector difference can be significant. It is quite probable that in these two documents, relative distributions of the terms may be equal; however, one may have a far larger absolute term frequency than the other. Computing the cosine similarity for the vectors representing these two documents is a method that can compensate for this document length effect.

Euclidean lengths apply to both documents and queries. The term *long document* is used to describe large documents where the possibility exists that two terms within a term dictionary exist far apart. A long document for example is typically a doctoral thesis consisting of 400 pages and a word count of 100,000 or larger, while a *short document* could be an abstract from a journal article with a word count of 200, similar to the document collection from the Cranfield collection setup in the 1950s (Cleverdon, 1956). As long documents can affect indexing granularity they could be subdivided into chapters (into pieces of shorter length) to make similarity comparisons more efficient (Hiemstra, 2000, 2009).

Cosine similarity theory encompasses numerous computations that help determine the similarity between two vectors in a vector space and assists the IRS to determine the comparative relevancy of them. The similarity can be between two documents vectors or between a query vector and a document vector. By definition, similarity is equal to the cosine of the angle between the two vectors. The angle between two document vectors  $d_3$  and  $d_4$  is represented by  $\text{sim}(d_3, d_4) = \cos \theta$  and between a

query vector  $q_1$  and a document vector  $d_4$  as  $\text{sim}(q_1, d_4) = \cos \theta$ . The vector representation for document  $d_3$  is  $\vec{V}(d_3)$  and for document  $d_4$  is  $\vec{V}(d_4)$ .

Utilising these vectors the cosine similarity (Equation 2.2) can be represented by

$$\text{sim}(d_3, d_4) = \cos \theta = \frac{\vec{V}(d_3) \cdot \vec{V}(d_4)}{|\vec{V}(d_3)| |\vec{V}(d_4)|}$$

**Equation 2.2: Cosine similarity sim (d3,d4)**

Referring to the similarity formula above, the numerator  $\vec{V}(d_3) \cdot \vec{V}(d_4)$  represents the dot product of the two vectors  $\vec{V}(d_3)$  and  $\vec{V}(d_4)$  while the denominator  $|\vec{V}(d_3)| |\vec{V}(d_4)|$  represents the product of the two document lengths.

Similarly, the vector representation for a query  $q_1$  is  $\vec{V}(q_1)$  and for document  $d_4$  is  $\vec{V}(d_4)$  and utilising these vectors the cosine similarity (Equation 2.3) can be represented, according to Manning et al. (2008) as:

$$\text{sim}(q_1, d_4) = \cos \theta = \frac{\vec{V}(q_1) \cdot \vec{V}(d_4)}{|\vec{V}(q_1)| |\vec{V}(d_4)|}$$

**Equation 2.3: Cosine similarity sim (q1,d4)**

Langville and Meyer (2007) however use a different notation for their formula (Equation 2.4), representing cosine similarity between a query vector  $q_1$  and a document vector  $d_4$ :

$$\delta_i = \cos \theta_i = \frac{q^T d_i}{\|q\|_2 \|d_i\|_2}$$

**Equation 2.4: Cosine similarity  $\delta_i$  (q1,d4)**

The cosine similarity for query vector  $q_1$  and a document vector  $d_4$  from the example (Equation 2.5) will therefore be:

$$\delta_4 = \cos \theta_4 = \frac{q_1^T d_4}{\|q_1\|_2 \|d_4\|_2}$$

**Equation 2.5: Cosine similarity  $\delta_4$  ( $q_1, d_4$ )**

As cosine similarity measures the similarity between a query and a document, it assists the IRS to determine the relevance between a query and a document. To assist in relevance a selected tolerance may be set to limit those documents retrieved as relevant. Berry et al. (1999) and Langville and Meyer (2007) argue that a selected tolerance can be used to specifically select a set level of tolerance in cosine similarity values, thus affecting those documents retrieved by the IRS.

Reflecting back to the vector space model example in the work of Langville and Meyer (2007), the authors compute the cosine similarity values for each of the seven documents and present their findings as:

$$\delta \approx [0 \ 0.40824 \ 0 \ 0.63245 \ 0.5 \ 0 \ 0.5]$$

In their work, Langville and Meyer (2007) use a selected tolerance of  $\tau = 0.1$  and therefore the retrieved documents returned to the user are only for documents where  $\delta_i > \tau$ . In this example, all four cosine similarity values of  $\delta$  that are greater than 0 are also greater than 0.1. However, if the selected tolerance was set to  $\tau = 0.45$  then this would only eliminate  $d_2 = 0.40824$ , leaving three cosine similarity values greater than 0.45 for  $d_4$ ,  $d_5$  and  $d_7$ .

## 2.10 Measurements and formulae

According to Manning et al. (2008), the core essence of an IRS is to address ad hoc retrieval tasks where documents within a collection are offered pertaining to an information need initiated by a user's query, which has been communicated to the IRS. The information need is not the query but the results stemming from it, and these may or may not satisfy the user's information need. An information need may encompass many queries each with its own query criteria. A document is judged relevant by the user only if the document offered (or retrieved) contains information that is relevant (has information value) to the users information need. This '*judgement*' by the user is time consuming; processing a query within an IRS may just take a few seconds but judging whether a document is relevant may take the user a few hours (Manning et al., 2008). Langville and Meyer (2007) define relevance as measure of similarity between a query and a document but does not necessarily satisfy a user's information need; a user must still intervene and judge whether the query and

document suggested by this similarity measurement is relevant. Manning et al. (2008) argue there are three items that are required when using an IRS and these are: i) a document collection; ii) a set of information needs that can be expressed as queries; and iii) a set of judgements expressed in binary to determine whether a query-document pair is either relevant or non-relevant. Judging whether a query-document pair is relevant or non-relevant is referred to as the *ground truth judgment of relevance* or the *gold standard of relevance*. The critical point of relevance is that it applies to an information need and not to a query, and hence the need for user intervention to determine whether a document is relevant or non-relevant. For testing purposes, at least fifty information needs are suggested (Manning et al., 2008).

An IRS can be viewed as a two-class classifier. By referring to Table 2.4, there are two classes deemed relevant and non-relevant. A two-class classifier is also referred to as a confusion matrix (Kohavi & Provost, 1998) or a 2x2 contingency table (Cleverdon & Keen, 1966; De Raadt et al., 2019). In this study, a two-class classifier will be referred to as a 2x2 contingency table. Table 2.5 illustrates an early example of a 2x2 contingency table by Cleverdon and Keen (1966) where each column represents *relevant* and *non-relevant* and each row as *retrieved* and *not retrieved*. Each of the rows  $a + b$  and  $c + d$ , and columns  $a + c$  and  $b + d$  are then summed with their totals. The total collection is then defined as  $a + b + c + d = N$ . Cleverdon and Keen (1966) citing numerous authors use specific formulae to define the various terms of recall<sup>45</sup>, snobbery ratio<sup>46</sup>, precision<sup>47</sup>, noise factor<sup>48</sup>, fallout<sup>49</sup> and specificity<sup>50</sup>.

Kohavi and Provost (1998) define the confusion matrix as a matrix that illustrates predicted and actual classifications, with L number of different label values, and with size  $L \times L$ . Table 2.4 represents a 2 x 2 confusion matrix where  $L = 2$ .

**Table 2.4: A confusion matrix (Kohavi & Provost, 1998:1)**

actual \ predicted	negative	positive
Negative	a	b
Positive	c	d

<sup>45</sup> Also referred to as sensitivity or hit rate

<sup>46</sup> The complement to recall

<sup>47</sup> Also referred to as relevance ratio, pertinency factor or acceptance rate

<sup>48</sup> The complement to precision

<sup>49</sup> Also referred to as fallout ratio

<sup>50</sup> The complement to fallout ratio

However in their work, Manning et al. (2008) refer to this matrix as a 2x2 contingency table with actual/predicted matrix inverted to outcome/actual thus rearranging the notations presented by Kohavi and Provost (1998), in alignment with the original work of Cleverdon and Keen (1966) and Cleverdon (1967), as illustrated in Table 2.5.

Table 2.5: A 2x2 contingency table (a,b,c,d) (Cleverdon & Keen, 1966:34)

	RELEVANT	NON-RELEVANT	
RETRIEVED	a	b	a + b
NOT RETRIEVED	c	d	c + d
	a + c	b + d	a + b + c + d = N (Total Collection)

Manning et al. (2008) further prefer to use the terms: true positive (*tp*), false positive (*fp*), true negative (*tn*) and false negative (*fn*) rather than the letters used by Kohavi and Provost (1998) (*d*, *c*, *b* & *a*) and refer to the columns as the truth (relevance) and the rows as the system (retrieved). The false positive concept originates from the *false drops* concept by Mooers (1950; 1951) – retrieved documents judged non-relevant. The ideas and concepts by Manning et al. (2008:143) presented in Table 2.6.

Table 2.6: A 2x2 contingency table (Manning et al., 2008:143)

	relevant	nonrelevant
retrieved	true positives ( <i>tp</i> )	false positives ( <i>fp</i> )
not retrieved	false negatives ( <i>fn</i> )	true negatives ( <i>tn</i> )

Reading Table 2.6 top-down and from a user's viewpoint, the number of relevant documents in a collection is represented by true positive (*tp*) plus false negative (*fn*) and the number of relevant documents is represented by true positive (*tp*). Reading this contingency table from left-to-right and from an IRS viewpoint, the number of documents that are retrieved by the system are represented by true positive (*tp*) plus false positive (*fp*) (Manning et al., 2008). Referring to Table 2.7, the number of relevant documents retrieved (*tp*) refers to the number of relevant documents judged relevant by the user, that were judged relevant and retrieved by the IRS; the number of documents retrieved (*tp + fp*) refers to the number of documents judged relevant and retrieved by the IRS; the number of relevant documents in collection (*tp + fn*) refers to all the relevant documents judged relevant by the user, from the entire document collection, whether the IRS retrieved them or not (Langville & Meyer, 2007; Manning et al., 2008). Referring to the formulae above defining both precision and recall, the formulae are now discussed.

### 2.10.1 Precision

Kent et al. (1955) introduced the concepts of precision and recall. Precision is the first of the two measurements that is utilised to measure an unranked IRS's effectiveness. In the work of Korde and Mahender (2012), the authors describe four IRS measurements: Precision, Recall, Fallout, and Accuracy. When discussing precision and recall, Weideman (2001) and Tonta (2019) describe what is not done as failures. A precision failure is the retrieval of non-relevant documents while a recall failure is a missed relevant document. According to Manning et al. (2008), to determine precision one must ask the question:

*“What fraction of the returned results is relevant to the information need?”*

Sanderson (2010) and Hardik and Jyoti (2012) describe precision as the number of retrieved documents that are relevant, and recall as the fraction of the total relevant documents retrieved. Hardik and Jyoti (2012) suggest that in recent times, mean average precision (MAP) values are now considered to give the best judgment when multiple queries are presented to the IRS.

Manning et al. (2008) define precision (P) as the fraction of retrieved documents from an IRS that are relevant (Equation 2.6) and is presented as:

$$Precision (P) = \frac{Relevant}{Retrieved} = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)}$$

Equation 2.6: Precision

Langville and Meyer (2007) however define precision (P) (Equation 2.7) as:

$$0 \leq Precision = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved}} \leq 1$$

Equation 2.7: Precision

Using the 2x2 contingency table and using the concepts *tp*, *fp*, *fn* and *tn*, Precision is defined as the ratio of the number of user relevant documents retrieved by the system and the total number of documents retrieved by the system (Manning et al., 2008). The equation (Equation 2.8) is thus presented as:

$$Precision (P) = \frac{tp}{tp + fp}$$

Equation 2.8: Precision

### 2.10.2 Recall

Recall is the second of the two measurements (Kent et al., 1955) that is utilised to measure an unranked IRSs effectiveness. According to Manning et al. (2008), to determine Recall, one must ask the question:

*“What fraction of the relevant documents in the collection was returned by the system?”*

Manning et al. (2008) define Recall (R) as the fraction of relevant documents from an IRS that is retrieved (Equation 2.9) and is presented as:

$$Recall (R) = \frac{Retrieved}{Relevant} = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)}$$

Equation 2.9: Recall

Langville and Meyer (2007) however define Recall (R) (Equation 2.10) in this way:

$$0 \leq Recall = \frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents in collection}} \leq 1$$

Equation 2.10: Recall

Using the 2x2 contingency table and using the concepts  $tp$ ,  $fp$ ,  $fn$  and  $tn$ , Recall is defined as ratio of the number of user relevant documents retrieved by the system and the number of user relevant documents in the collection. The equation (Equation 2.11) is presented as:

$$Recall (R) = \frac{tp}{tp + fn}$$

Equation 2.11: Recall

### 2.10.3 F-measure

The way to access the combined measure between the trade-off of Precision (P) and Recall (R), is by using the weighted harmonic mean known as the F-measure. Origins

of the F-measure are traced back to the work of van Rijsbergen (1979) where the author makes reference to the complement of the F-measure, which is stated as  $E = 1 - F$ . It is in this work that van Rijsbergen (1979) provides reasoning for the adoption of the harmonic mean as a method when combining the values of both precision and recall.

The F-measure (Equation 2.12) is typically presented as:

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Equation 2.12: F

The most common measure is the balanced  $F_1$ -measure where beta  $\beta = 1$  and alpha  $\alpha = \frac{1}{2}$ . By substituting these values into the formula, F (Equation 2.13) now becomes:

$$F_1 = \frac{2PR}{(P + R)} \text{ where } \beta = 1, \alpha = \frac{1}{2}$$

Equation 2.13: F1

F-measure is a measure of the overall effectiveness of an information retrieval system, and uses the product and sum values of Precision and Recall. The equation (Equation 2.14) is presented as:

$$F - \text{measure} = \frac{2PR}{(P + R)}$$

Equation 2.14: F-measure

#### 2.10.4 Accuracy

Accuracy attempts to measure that fraction of classifications the system considers to be correct. An IRS retrieves documents (offers documents to the user) it considers relevant. An alternative measure that can inform a user when making a judgement of relevancy is accuracy (Zhao, 2012). In their work, Kohavi and Provost (1998) define accuracy (Equation 2.15) as:



$$Accuracy(A) = \frac{a + d}{a + b + c + d}$$

Equation 2.15: Accuracy (a,b,c,d)

Similarly, Manning et al. (2008) define accuracy (Equation 2.16) as:

$$Accuracy (A) = \frac{(tp + tn)}{(tp + fp + fn + tn)}$$

Equation 2.16: Accuracy (tp,fp,fn,tn)

### 2.10.5 Snobbery ratio

Snobbery or snobbery ratio is the complement of recall: the ratio of the number of relevant documents not retrieved and the number of relevant documents in a collection. Snobbery (Equation 2.17) is defined by Cleverdon and Keen (1966) as:

$$Snobbery (Sn) = \frac{fn}{tp + fn}$$

Equation 2.17: Snobbery ratio

### 2.10.6 Noise factor

Noise factor is an undesirable or unwanted input and Hauff (2010) suggests that noise factor should be considered when evaluating IRSs.

The formula for noise factor (Equation 2.18) is presented as:

$$Noise\ factor\ (Nf) = \frac{fp}{tp + fp}$$

Equation 2.18: Noise factor

### 2.10.7 Fallout

Weideman (2001), Korde and Mahender (2012), and Tonta (2019) discuss the use of measuring fallout in IRSs and especially the failures. Weideman (2001) argues that fallout is a failure, as it is the retrieval of too many irrelevant documents. Egghe (2008), Sanderson (2010), and Hardik and Jyoti (2012) describe fallout as a measurement of the fraction of non-relevant documents retrieved. Sanderson (2010) states that although fallout is often described in the IRS text books, it is the measurement least

used in published IRS research. The formula for fallout (Equation 2.19) is presented as:

$$Fallout (Fo) = \frac{fp}{fp + tn}$$

Equation 2.19: Fallout

### 2.10.8 Specificity

Based on the work of Cleverdon and Keen (1966) using phrase based indexing, Croft et al. (1991) argue that, if used correctly phrase-terms and the indexing thereof, should improve the specificity of an IRS. In their work, Dinh and Tamine (2015) used average specificity and calculated the overall average specificity for each of their queries. In healthcare, Chunara, Freifeld and Brownstein (2012) used ‘sensitivity’ and ‘specificity’ while Choudhary et al. (2017) used the mean of ‘average specificity’. Dinh and Tamine (2015) evidenced that the more the query terms were specific, the more the specificity of the returned documents was finely grained. Spärck Jones (1972) suggests that in IRS evaluation, specificity should be interpreted statistically, as a function of term use rather than of term meaning. The formula for Specificity (Equation 2.20) is presented as:

$$Specificity (S) = \frac{tn}{fp + tn}$$

Equation 2.20: Specificity

In IRSs, Spärck Jones (1972:2) refers to measuring specificity as “statistical specificity”. In open documents collections such as the Web ‘tn’ cannot be quantified as the number of documents in the collection (N) is unknown and  $N = tp+fp+fn+tn$ . However, in closed collections, N is known, and therefore ‘tn’ can be calculated and therefore specificity can be measured. Henderson (1967:119) support this and states: “Specificity takes into account one of the vital parameters in a retrieval system ... size of file”.

### 2.10.9 Measurements of agreement

To test the strength of agreement between the judgments made by the IRS and those judgements made by the user, the Kappa coefficient is one method based on the work of Cohen (1960). The Kappa coefficient measures the consistency of agreement between two judgements (Cohen, 1960). Conger (2017) and de Raadt et al. (2019) concur that it is a coefficient commonly used for measuring the degree of agreement between two measures on a nominal scale. Proposed scales for measurements of

agreements have been proposed by Landis and Koch (1977) and Fleiss, Levin and Paik (2003).

The origins of judging relevancy using the Kappa coefficient introduced by Cohen (1960) stems from his work entitled, '*A coefficient of agreement for nominal scales*'. The Kappa coefficient, sometimes referred to as the Kappa statistic (Manning et al. 2008), measures agreements between judges, and in the case of information retrieval often between system judgment and user judgement. It is designed to provide the definitive judgment (Cohen, 1960). The following formulae (Equation 2.21) can now be used to calculate the Kappa coefficient beginning with the observed proportion of the number of times both judges agreed:

$$P(A) = \frac{tp + tn}{N}$$

Equation 2.21: P(A)

Then the pooled marginal of P(nonrelevant) and P(relevant), as

$$P(\text{nonrelevant}) = \frac{(fn + tn) + (fp + tp)}{N + N}$$

Equation 2.22: P(non-relevant)

$$P(\text{relevant}) = \frac{(tp + fp) + (tn + fn)}{N + N}$$

Equation 2.23: P(relevant)

Then the probability both judges agreed by chance, as

$$P(E) = P(\text{nonrelevant})^2 + P(\text{relevant})^2$$

Equation 2.24: P(E)

Then finally the Kappa coefficient (Equation 2.25), is

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

Equation 2.25: k

#### 2.10.10 Scale of agreed judgments

To determine the matching of judgements, the strength of agreement between the judgments made by the IRS and the user, a six-division range has been proposed by Landis and Koch (1977). This range is from 'poor' to 'almost perfect', indicated by values ranging from less than 0 to 1. These six divisions of the measurement range for the Kappa coefficient are presented as:  $k < 0.00$  as poor,  $0.00 \leq k \leq 0.20$  as slight,  $0.21 \leq k \leq 0.40$  as fair,  $0.41 \leq k \leq 0.60$  as moderate,  $0.61 \leq k \leq 0.80$  as substantial, and  $0.81 \leq k \leq 1.00$  as almost perfect.

Based on the work of Landis and Koch (1977), Fleiss et al. (2003) argue for a three-division range, and suggest:  $k > 0.75$  as excellent,  $k < 0.40$  as poor, and a  $k$  value between 0.40 and 0.75 as fair to good agreement beyond chance. However, Manning et al. (2008) argue a different scale for measuring the strength of agreement between the IRS and the user and state that if  $k = 1$  the two judges always agree, if  $k > 0.8$  then the agreement between the two judges is good, if  $0.67 \leq k \leq 0.8$  then the agreement is fair, if  $0 \leq k < 0.67$  the agreement is doubtful, and if  $k$  is negative then the chance the judges will agree is worse than random.

#### 2.10.11 IRS significance tests

In their work, Smucker, Allan and Carterette (2007) have performed a comparison of statistical significance tests for the evaluation of IRSs. The authors compare the Wilcoxon signed-rank test (Maddalena et al., 2017), the sign test (Trieschnigg, 2010; Maxwell, 2014), the student paired t-test (Trieschnigg, 2010), the bootstrap (Shekarpour et al., 2017) and the randomisation test of Fisher (1971). In performing this comparison, Smucker et al. (2007) use mean average precision (MAP) to statically test for significance.

Of the five tests used, their results suggest that the Wilcoxon and sign tests are simplified variants of the randomisation and should not be used in IRS significance testing as they have poor ability. However, of the remaining three, little practical difference is found between them and suggests that the student paired t-test, the bootstrap, and Fisher's randomisation (permutation) test should be used in comparing IRSs (Smucker et al., 2007).

## 2.11 The theoretical conceptual framework

After a review of the literature, the stages and key concepts of information retrieval derived from the literature are presented. Taking cognisance of the two research questions and the five hypotheses for this research, the theoretical conceptual framework for this study is presented.

General systems theory is used as the theoretical lens for this study, based on the work of Von Bertalanffy (1968:162) where a simple feedback scheme is used. This scheme (Figure 2.11) is based on the classical stimulus-response scheme where the feedback loop is added providing a circular interconnection.

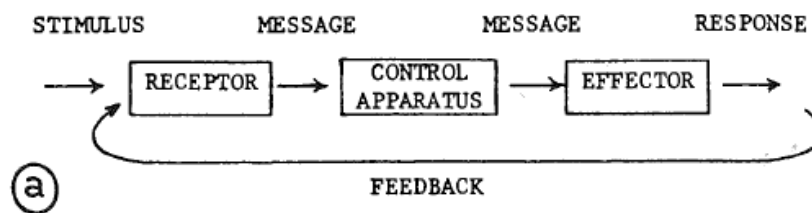


Figure 2.11: A simple feedback scheme (Von Bertalanffy, 1968:162)

The key concepts for each of these three stages, derived from the literature review, are:

- i) The user stage (receptor) pertains to the manual activities made by the user: the first is the collection of documents to be analysed, the second is the information need of the user, and the third is the judgement made by the user – whether each document in the collection satisfies (or not) each need of the user.
- ii) The IRS stage (control apparatus) pertains to the programmatic activities made by the IRS: firstly information from the documents is gathered and thereafter this information is stored in an index, secondly the search engine is activated to search for terms in a query that match words in a document, thirdly is the judgment made by the IRS – if the query term matches a token in the index then the relevant document(s) is/are retrieved, and fourthly, does the IRS solve the problem of vocabulary mismatch?
- iii) The evaluation stage (effector) takes the judgement results made by the user and by the IRS and these results are statistically analysed in order to measure the performance of the IRS and the performance of the user.

Table 2.7 presents the three stages, together with the key concepts and the sections that pertain to them, illustrated in the theoretical conceptual framework derived from the literature.

Table 2.7: The stages and key concepts derived from the literature

Stage	Concept 1	Concept 2	Derived from
User	Documents in a collection		Section 2.2: A brief history of information retrieval Section 2.4: A user's pursuit for documents Section 2.5: IRS models and methods Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.9: IRS evaluation Section 2.10: Measurements and formulae
	Information need		Section 2.2: A brief history of information retrieval Section 2.4: A user's pursuit for documents Section 2.5: IRS models and methods Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries
	User judgement	<i>tpfn</i> relevant <i>fptn</i> non-relevant	Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.10: Measurements and formulae
IRS	Information gathering	Content acquisition	Section 2.7: IRS design concepts
		Text transformation	Section 2.7: IRS design concepts
		Token index	Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts
	Search engine	Term	Section 2.2: A brief history of information retrieval Section 2.4: A user's pursuit for documents Section 2.5: IRS models and methods Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.9: IRS evaluation Section 2.10: Measurements and formulae
		Query	Section 2.2: A brief history of information retrieval Section 2.4: A user's pursuit for documents Section 2.5: IRS models and methods Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.9: IRS evaluation Section 2.10: Measurements and formulae
	IRS judgement	<i>tpfp</i> retrieved	Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.10: Measurements and formulae
<i>fnfn</i> not retrieved		Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.10: Measurements and formulae	
Vocabulary mismatch		Section 2.3: Concepts of vocabulary mismatch Section 2.8: Search engines and queries	
Evaluation	Performance measurements	<i>tp, fp, fn, tn</i>	Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.10: Measurements and formulae
		Precision	Section 2.2: A brief history of information retrieval Section 2.5: IRS models and methods Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.9: IRS evaluation Section 2.10: Measurements and formulae
		Recall	Section 2.2: A brief history of information retrieval Section 2.5: IRS models and methods Section 2.6: IRS Indexing methods Section 2.7: IRS design concepts Section 2.8: Search engines and queries Section 2.9: IRS evaluation Section 2.10: Measurements and formulae

Stage	Concept 1	Concept 2	Derived from
		F-measure	Section 2.2: A brief history of information retrieval Section 2.5: IRS models and methods Section 2.6: IRS Indexing methods Section 2.8: Search engines and queries Section 2.9: IRS evaluation Section 2.10: Measurements and formulae
		One-tailed student's t-test	Section 2.10: Measurements and formulae
		Kappa coefficient	Section 2.2: A brief history of information retrieval Section 2.10: Measurements and formulae

The theoretical conceptual framework for this study, indicating the scope of the study and the main elements within it, is presented. The framework offers a design strategy of how to design an IRS and once built, of how to use it and evaluate it. For clarity and providing context to the framework, the research questions and five hypotheses are re-stated:

**RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

**H1<sub>0</sub>:** Hybridised indexing does not increase the effectiveness of retrieving relevant documents

**H2<sub>0</sub>:** Hybridised indexing does not reduce the incorrect identification of relevant documents

**H3<sub>0</sub>:** Hybridised indexing does not increase the quality in rejecting non-relevant documents

**H4<sub>0</sub>:** Judgments made by the hybrid indexing method and the user disagree

**H5<sub>0</sub>:** The hybrid indexing method does not satisfy the information needs of the user

**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

Referring to the theoretical conceptual framework (Figure 2.12), the three stages are presented: the user stage, the IRS stage, and the evaluation stage.

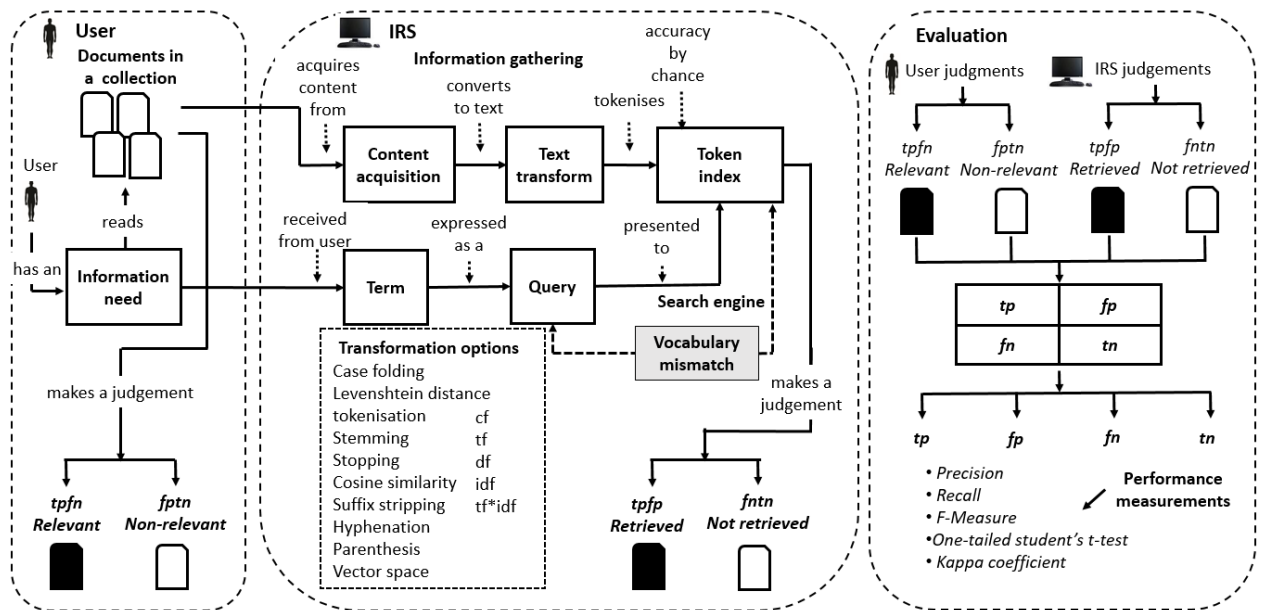


Figure 2.12: A theoretical conceptual framework from the literature

### 2.11.1 The user stage

The user stage begins when a user (a researcher) has an information need that he/she desires to be satisfied. The user then gathers all his/her documents collected over the years pertaining to the research topic and then reads through each document individually in the language of his/her choice. While perusing each document within this collection, at some point the user makes a judgement about whether the document is relevant (*tpfn*) or non-relevant (*fptn*) to his/her information need.

### 2.11.2 The IRS stage

The IRS stage tries to mirror those processes made by the user in the user stage using a computer. There are two main processes during the IRS stage: the information gathering process and the search engine process. During the information gathering process the IRS acquires the content from each document by firstly using OCR software (Waitelonis, 2018) to convert the source file format to text file format; secondly by transforming the text by replacing a few or all special characters (hyphens, punctuation, parenthesis, delimiter) (Hudson & Manning, 2019) and then choosing whether to simultaneously case fold all the text to lowercase (Agnihotri et al., 2017); thirdly by tokenisation where the chunks of text between the delimiters are acquired and stored as tokens, together with their document identifiers within the token index. Token index design differs greatly and there are many suggestions from the literature, each with their own distinguishing properties (Manning et al., 2017; Liao et al., 2019). During the search engine process, terms (words) are received from the user and these terms are then expressed as queries, which are a representation of the users information need. Once the query is presented to the search engine,



attempts are made to match each term within the query to the tokens that exist within the index – this is where the problem of vocabulary mismatch occurs. The IRS then makes a judgement as to whether to retrieve (*tpfp*) or not retrieve (*fnfn*) the document from the collection. During this judgment process there are a number of ranking and weighting methods that can be utilised in order to aid the IRSs judgement decision, for example, stopping (Hudson & Manning, 2019), stemming (Bi, Ai & Croft, 2019), cosine similarity (Orkphol & Yang, 2019), vectors (Agnihotri et al., 2017), *idf* (Waitelonis, 2018), and *tf\*idf* (Skrlj et al., 2019). These methods are what make search engines differ in their judgement performances.

### 2.11.3 The evaluation stage

The evaluation stage works with the judgement results made by the user (*tpfn* and *fpfn*) and the results made by the IRS (*tpfp* and *fnfn*). These results are dropped into a 2x2 contingency table (De Raadt et al., 2019) and from this table the values of *tp*, *fp*, *fn* and *tn* are derived. To measure the performance of the IRS (that is how good the computer's judgement is compared to the users judgement) various formulae are available. The most commonly used formulae according to the literature in IRS evaluation appear to be Precision, Recall, and F-measure (Narayan et al., 2017). To compare two systems statistically, the one-tailed student's t-test (Huang & Huang, 2016) recommended from the investigation performed by Smucker et al. (2007) and the Kappa coefficient (Conger, 2017) introduced by Cohen (1960) nearly 60 years ago, are used.

In section 3.1, the problem statement, the research questions, and five hypotheses are re-stated to provide further context, and in Figure 3.1 an updated framework is presented that positions the research questions and hypotheses within the framework's main elements.

## 2.12 Summary

The intention of section 2.2, '*A brief history of information retrieval*', was to research the key role players in information retrieval, what they did, how they did it, and to determine the contributions made by them. Information retrieval began many years before the advent of the computer and as such, the goal was to determine the manual methods used as these could be programmatically simulated if required. The review was successful as it provided good insight into how IRSs evolved over the last century and how attempts were made to solve many of the challenges encountered successfully. However, what was lacking in more recent times were the actions of the user and how the user actually performed his/her judgements when IRSs needed to be evaluated. This formed the basis for the next section.

Section 2.3, '*Concepts of vocabulary mismatch*', was an investigation into the problem of vocabulary mismatch, to define what it actually is, how it occurs, and how it was discovered. The review was successful as it provided the insight into the origin of the problem and how and why the problem still exists today in information retrieval.

The intention of section 2.4, '*A user's pursuit for documents*', was to dig deeper into what was designed in the past, how these manual processes worked and what were the problems that were overcome to achieve the needs of the user in pursuit of documents. This review was partially successful; the manual methods were well described but the more recent programmatic methods were not. Therefore, further investigation into more recent models and methods was needed.

Section 2.5, '*IRS models and methods*', was an investigation into the various IRS models and methods. This began with the bag of words model that explained the basic principles of word ordinality and proximity, the citation ranking method, the Boolean retrieval model, the Vector space model, the Markov random field model, and the contributions from Zipf's law. This review was successful as it described the various approaches many researchers had in attempting to solve problems and how mathematics was introduced to assist with providing solutions. However, the key concept to storing document information was the index, and since these models and methods were not index focused, further investigation into the technical design of these indices was needed.

Therefore, the goal of section 2.6, '*IRS Indexing methods*', was to determine the design concepts of these indices because it was possible that the root cause of mismatching words between a query and a document lay in the design of the index. Thus the research questions are stated:

**RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

Although many indices were reviewed, none of those reviewed had the ability to solve the vocabulary mismatch problem. Nevertheless, there was a chance that by mixing concepts from one or more indices used in information retrieval and by using the ideas of data retrieval discussed in section 2.4, a new indexing design could be presented and evaluated. Although the review of indices was comprehensive, and one of the

objectives of this research was to design and build a new indexing method, putting theory into practice was the difficult challenge. Although indices were reviewed, there was a need to investigate the concepts in the design of the whole IRS.

The first objective of this research was to design, build, and rigorously pilot test a new indexing method and therefore the intention of section 2.7, '*IRS design concepts*', was to dig deeper into IRS design and to determine what and how it has been done. There were three parts to this: acquiring the contents of documents as text, transformation the text, and storing the data in a token index. The literature is replete with text transformation processes and it appears this is where many IRS researchers focus their research, for example, case folding, tokenisation, stemming, stopping, cosine similarity, suffix stripping, hyphenation, and vectors. As many of these processes influence, through weightings, the results of the IRS the focus of this research was to design and build a new IRS, with a novel indexing method using non influential concepts such as stopping and stemming, and to rigorously pilot test it. Thereafter it would need to be evaluated to determine whether or not it was more effective than one of the traditional indexing methods. Now that the process of gathering the information from documents through content acquisition, text transformation, and token index population was understood, an investigation into the search engine query process was needed.

The intention of section 2.8, '*Search engines and queries*', was therefore to understand how queries are expressed and how they are presented to the search engine. The two main concepts were the *term* and the *query*, where one or more terms are expressed within a query and presented to the search engine to perform the search. Here lay the root cause of the problem of vocabulary mismatch: the matching of the term in a query to a term within a document. Queries can be expanded so that they include multiple terms but the use of phrase-terms was seldom found in the literature. The concepts of word ordinality and proximity were discussed but were not ideally handled within the functionality of the indexing systems reviewed. At this stage of the literature review, a gap appeared between the design of the indices and how the search engine interacted with these indices to retrieve relevant documents, while at the same time attempting to solve the problem of mismatched vocabulary. This is when the idea of designing a hybrid indexing method occurred to this researcher and the need to test it by building a complete IRS, but a further literature review was required to determine what methods have already been used to evaluate an IRS.

Therefore, section 2.9, '*IRS matching and processing*', describes how an IRS is evaluated and how the mathematical concepts are used in IRS evaluation. At this stage after examining the concepts, it was not known which would be used in this research, so a full investigation of the literature was needed. This was a challenging review as each of these mathematical concepts had to be practically tested using programmatic prototypes to fully understand how they worked and how they interacted with each other. In addition if any of these concepts from the literature were to be used in this research the software adopted had to be capable of handling these concepts. Although frequency calculations were used effortlessly large multiple column term-by-document matrices became a challenge. Again, a gap appeared in the literature because there was no full practical description of how these mathematical concepts fit in to the design of an IRS. However once the mathematical concepts for IRS evaluation were explored the final requirement was to investigate how these data should be presented to test the three hypotheses.

The final section 2.10, '*Measurements and formulae*', investigated the literature to determine which measurements and formulae were appropriate to test the three hypotheses:

**H1<sub>0</sub>**: Hybridised indexing does not increase the effectiveness of retrieving relevant documents

**H2<sub>0</sub>**: Hybridised indexing does not reduce the incorrect identification of relevant documents

**H3<sub>0</sub>**: Hybridised indexing does not increase the quality in rejecting non-relevant documents

The three popular formulae to evaluate an IRS were Precision, Recall, and F-measure. These were explored in section 2.3 and in further detail in this section 2.10, with eight formulae being derived from the 2x2 contingency table. The remaining five of the eight formulae for the evaluation criteria were sought and found in the literature, with difficulty, and these were: Accuracy, Snobbery ratio, Noise factor, Fallout and Specificity. This review (including section 1.7) was successful as Precision (together with mean Precision and MAP) was applied to **H1<sub>0</sub>**, Recall (together with MAR) was applied to **H2<sub>0</sub>** and Specificity (together with MAS) was applied to **H3<sub>0</sub>**. The final two hypotheses are:

**H4<sub>0</sub>**: Judgments made by the hybrid indexing method and the user disagree

**H5<sub>0</sub>:** The hybrid indexing method does not satisfy the information needs of the user

To test these hypotheses, a further review was required to determine how to measure two sets of IRS judgments and how to scale those measurements, which revealed the Kappa coefficient and other scales suggested from by other authors. This part of the review was also successful as those measurements and formulae reviewed, and thereafter those selected, were confirmed by a professionally registered statistician as the appropriate measurements and formulae to use for these hypotheses.

From the literature, numerous IRS models and indexing methods were explored, and their benefits identified, together with mathematical techniques that could enhance or suppress certain features of indexing, and their performance results. Theories and techniques that suggested word ordinality and word proximity could be maintained, how the vocabulary mismatch problem could occur, and what methods could be used to satisfy a user's information needs, were explored. Further, suggested IRS performance measurements were determined including those frequently used (more Web focused) and those infrequently used (more stand-alone document collection focused) measurements. The gap found in the literature was twofold: firstly there was no indexing method that could maintain the sequencing of words in the correct order, that could exactly match two or more word phrase-terms expressed as a query, and hence a new indexing method needed to be designed, build and proved to work; and secondly, there was no technique available to measure these phrase-terms for the new indexing method.

In summary, the literature review has presented the key concepts and techniques that apply to IRSs, the various information retrieval models, the process of information gathering and the population of an index, the search engine process and query formulation, term weighting, term proximity, term frequencies and ranking, and finally, evaluation matrices and their many formulae. The chapter concludes by providing the design view of the theoretical conceptual framework, using terms and concepts from the literature, for this new indexing method.

Chapter Three provides the research methodology including the research purpose, philosophy, and strategy needed to address the research problem.

### 3. CHAPTER THREE: RESEARCH DESIGN

*"It is better to do the right thing wrong than to keep doing the wrong thing better and better!"*

– Russell Achoff

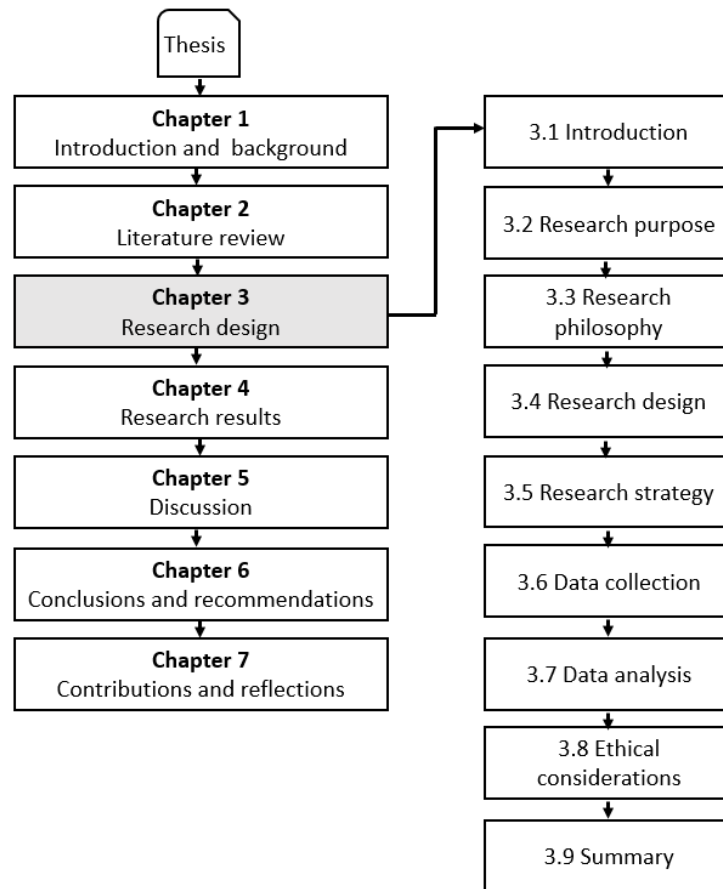


Figure 3.1: Schematic representation of Chapter Three

#### 3.1 Introduction

The chapter begins with a re-statement and an overview of the research problem, the research questions and the five hypotheses are presented, and the research purpose is explained. Thereafter the research philosophy is presented, followed by the research strategy, data collection, data analysis, and ethical considerations.

The problem statement for this study is as follows:

Challenges exist for information retrieval systems in handling mismatching vocabularies in queries and candidate source documents (Onal et al., 2018). As a result, these information retrieval systems may retrieve some documents that are non-relevant and miss some that are relevant (Van Gysel, 2017). This increases the time for research by forcing additional perusal of unsatisfactory results, and additional searches using alternative vocabularies (Liu et al., 2017). This renders information

retrieval systems less effective than they could be, and inhibits productive research (Mitra & Awekar, 2017; Nguyen et al., 2018).

With the problem stated, the first research question relevant to the design and build of the IRSs is presented. Next, the three hypotheses that were used to prove that the hybrid indexing method works are presented, followed by the last two hypotheses that are designed to find the answers to the problem statement. Finally, the second research question, which summarises this research, is presented.

**RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

The aim of this research question was to design, build, and rigorously pilot test a hybrid indexing method (IRS-H) that maintains word ordinality and word proximity, and to compare the effectiveness of this method with the traditional inverted indexing method (IRS-I). The research method used a literature review; the design and build of both IRSs followed by three pilot tests, and an experiment with users completing a questionnaire. Next, an evaluation was performed between IRS-H and IRS-I, where after the performance measurements were calculated. Referring to the updated conceptual framework that now illustrates the positioning of the first research question and hypotheses in Figure 3.2, RQ1 falls within the IRS stage, with specific emphasis on the design of the token index – this is where the search engine query attempts to match its query terms with the tokens stored in the token index.

**H1<sub>0</sub>:** Hybridised indexing does not increase the effectiveness of retrieving relevant documents

The objective of the first hypothesis was to test whether an IRS using a hybrid indexing method increases the effectiveness of retrieving only those documents that are judged relevant by the user. The research method used a literature review, the performance measurements generated by the IRSs and the users, the precision formula, an average precision ranking method, the MAP formula, and a statistical analysis using a one-tailed student's t-test. Referring to the updated conceptual framework, **H1<sub>0</sub>** falls within the Evaluation stage and requires the user-generated value of *tpfn* from the user stage and the IRS generated value of *tpfp* from the IRS stage to derive *tp* and thereafter Precision, where,  $P = tp / (tp + fp)$ .

**H2<sub>0</sub>:** Hybridised indexing does not reduce the incorrect identification of relevant documents

The objective of the second hypothesis was to test whether the hybrid indexing method reduces errors in incorrect identification of user judged relevant documents thus reducing the number of documents for the user to peruse. The research method used a literature review, the performance measurements generated by the IRSs and the users, the Recall formula, an average recall ranking method, the MAR formula, and a statistical analysis using a one-tailed student's t-test. Referring to the conceptual framework, **H2<sub>0</sub>** falls within the Evaluation stage and requires the user-generated value of *tpfn* from and the IRS generated value of *tpfp* to derive *tp* and thereafter Recall where,  $R = tp/(tp + fn)$ .

**H3<sub>0</sub>**: Hybridised indexing does not increase the quality in rejecting non-relevant documents

The objective of the third hypothesis was to test whether the hybrid indexing method increases the rejection quality of user non-relevant documents, thus providing confidence to the user in the judgement of the IRS. The research method included a literature review, the performance measurements generated by the IRSs and users, the specificity formula, an average specificity ranking method, the MAS formula, and a statistical analysis using a one-tailed student's t-test. Referring to the conceptual framework, **H3<sub>0</sub>** falls within the Evaluation stage and requires the user-generated value of *fptn* from the user stage and the IRS generated value of *fnfn* from the IRS stage to derive *tn* and thereafter Specificity, where,  $S = tn/(fp + tn)$ .

**H4<sub>0</sub>**: Judgments made by the hybrid indexing method and the user disagree

The objective of the fourth hypothesis was to determine whether the judgments made by the hybrid indexing method and the user agree. The research method used a literature review, the user judgements acquired from the questionnaire, the judgement results generated by IRS-H and IRS-I, and the calculation of the Kappa coefficient producing the agreement measurements. Referring to the conceptual framework, **H4<sub>0</sub>** falls within the IRS stage to determine whether the terms within the queries match the tokens stored within the token index.

**H5<sub>0</sub>**: The hybrid indexing method does not satisfy the information needs of the user

The objective of the fifth hypothesis was to determine whether the hybrid indexing method satisfies the information needs of the user by retrieving those documents from the collection that are relevant to the user. The research method used a literature review, the user judgements acquired from the questionnaire, the judgement results



generated by IRS-H and IRS-I, and the calculation of the Kappa coefficient producing the agreement measurements. Referring to the conceptual framework, **H5<sub>0</sub>** falls within the User stage but draws on the IRS generated data of *tpfp* to determine whether the documents retrieved by the IRS (*tpfp*) match those judged relevant by the user (*tpfn*).

**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

The objective of the second research question was to determine whether the hybrid indexing method solved the problem of mismatched vocabulary between a query and a document. This question is answered after a thorough literature review, after the exploratory and explanatory studies are completed, after the results from the first research question are presented, and after the five hypotheses and research findings are concluded. Referring to the updated conceptual framework that now illustrates the positioning of the second research question, RQ2 falls within the IRS stage positioned between the query and the token index. This is where the problem of vocabulary mismatch occurs. Figure 3.2 summarises a complex procedure, which is first presented as a narrative summary before a more detailed explanation in the paragraphs that follow.

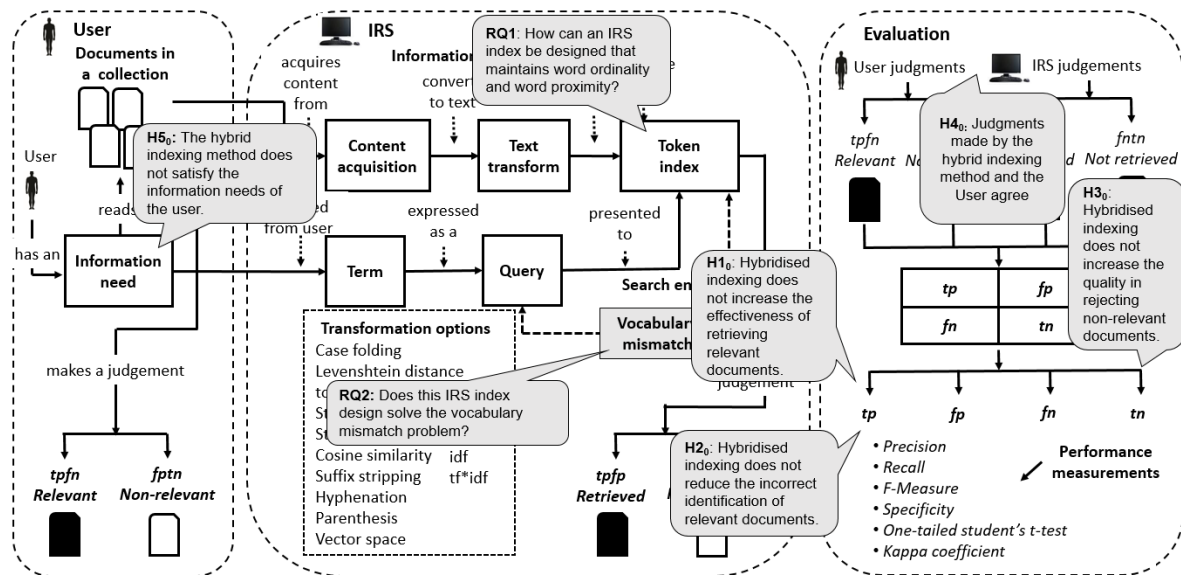


Figure 3.2: Theoretical conceptual framework with research question and hypotheses

### 3.2 Research purpose

According to Saunders, Lewis and Thornhill (2009) there are three classifications by which one can define the research namely: i) the exploratory study; ii) the descriptive study; and iii) the explanatory study.

### 3.2.1 The exploratory study

Robson (2005) suggests that the exploratory study has the ability to determine what is actually happening when phenomena are approached with differing views, when the necessary questions are asked, and by accessing the phenomena differently. Saunders et al. (2016) argue that there are three methods of performing an exploratory study, the first is by conducting interviews with the experts, the second by conducting interviews with specialist focus groups, and the third by performing a literature analysis. Babbie (2013) suggests that an exploratory study, if well done, can help focus future research while Saunders et al. (2019) warn that the researcher must be willing to redesign, change, or redirect what is being done when new information or fresh ideas become available.

In this study, the literature was reviewed numerous times as depicted in the flow chart (Figure 3.3). The purpose of this research was to make this activity of reviewing the literature numerous times more effective and to reduce the time taken in performing this. This is explained in the flow chart that represents how this exploratory study unfolded.

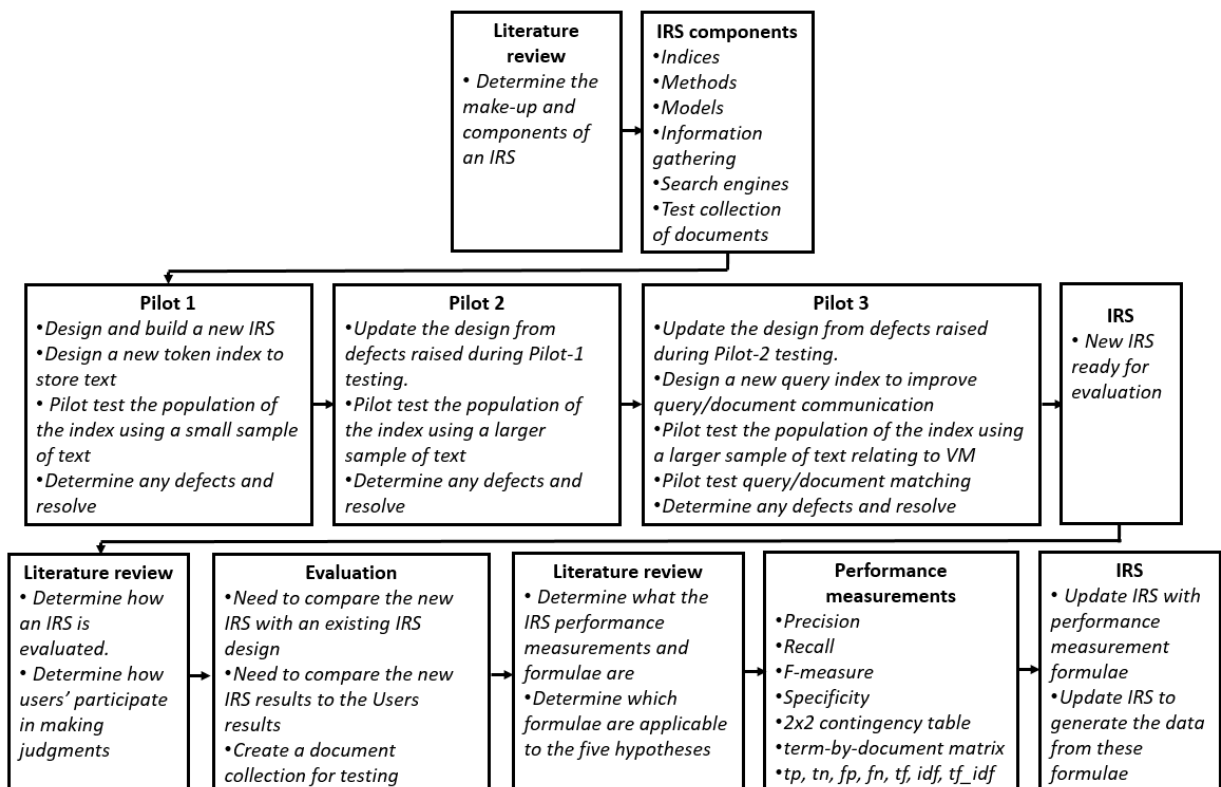


Figure 3.3: A flow chart representing the exploratory study

Reading from left-to-right, the literature was revisited to determine how an IRS is made-up and what the key components are. In summary, these were the various indices, IR models and methods, the gathering of information from text to index, how

search engines worked and the need for a test collection of documents. Using a few of these key components a new token index for storing text was designed and integrated into a newly designed and built IRS.

Pilot testing began using a small sample of text, the results were analysed, and defects were raised if there were any design issues (Volume II, Appendix A for Pilot 1 testing results). Once updates to the IRS and index design were completed, a second cycle of pilot testing was performed by populating the token index with a larger sample of text and defects were raised where necessary (Volume II, Appendix B for Pilot 2 testing results).

Updates from Pilot 2 defects were performed to the IRS and index design. A new query index was designed and integrated into the IRS to perform query searches used by the search engine.

Thereafter, a third cycle of pilot testing was performed by populating the token index with an even larger sample of text. This text related to vocabulary mismatch articles and was used to test the query document matching used by the search engine. Finally, defects were raised where necessary (Volume II, Appendix C for Pilot 3 testing results) and the new IRS was now ready for evaluation.

At this stage of the research the literature was revisited to determine how an IRS is evaluated, and how users participate in making their judgements of whether documents are relevant or not. To prove one method was more effective than another, there was a need to compare the new IRS with an existing IRS design and to compare results from the IRS with those generated by the user (section 4.9). The final requirement for evaluation was creating a collection of documents for testing purposes.

A further search of the literature was performed to determine what performance measurements are used by an IRS, what the formulae are associated with these measurements, and how these formulae can be used to test the five hypotheses. The performance measurements identified for this study were Precision, Recall, F-measure, Specificity, the 2x2 contingency table, the term-by-document matrix, and  $tp$ ,  $tn$ ,  $fp$ ,  $fn$ ,  $tf$ ,  $idf$  and  $tf\_idf$ . The final stage of this exploratory study included updating the IRS with these formulae to generate decision data using these performance measurements and formulae.

### 3.2.2 The descriptive study

Robson (2005) argues that the objective of a descriptive study is to describe accurately the events, situations, or profiles of people. Saunders et al. (2016) concur and suggest this can be achieved when a comprehensive view of the phenomenon is established, and it could be an extension to exploratory or explanatory studies.

A descriptive study was not considered for this research as it is based on describing characteristics of a population and is traditionally used to answer the 'what' question and not the 'how' question and the testing of hypotheses as in this research.

### 3.2.3 The explanatory study

Babbie (2013) suggests that an explanatory study provides reasons for phenomena in the form of causal relationships. Saunders et al. (2016) concur by describing an explanatory study as one where a problem is identified and is then best described by establishing the causal relationships between variables. This could be achieved quantitatively using statistical correlations based upon the data collected. Robson (2005) does warn the researcher that the research purpose may change during the study and may have to be adapted for the changed purpose.

The flow chart in Figure 3.4 represents how the explanatory study for this research, in the form of an experiment, was prepared for and how it was conducted.

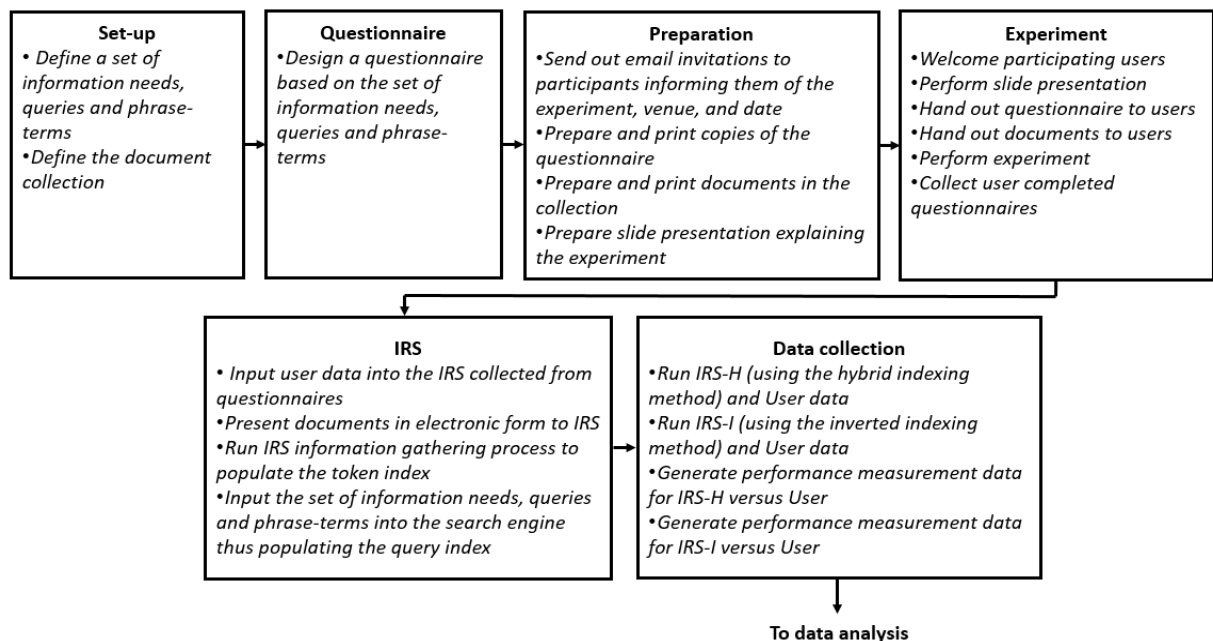


Figure 3.4: A flow chart representing the explanatory study

Setting up the data for the experiment included defining the definitive set of information needs, queries and phrase-terms together with the closed collection of

documents. Thereafter, and based on these information needs, queries and phrase-terms, the questionnaire was designed. In preparation for the experiment, email invitations were sent to the participants informing them of the experiment, venue, and date. Thereafter, copies of the questionnaire and the documents in the collection were printed and a slide presentation explaining the experiment was prepared. On the day of the experiment, the participating users were welcomed, followed by an introductory slide presentation of the experimental process. The questionnaires were handed out to the users and the users selected their set of printed documents. The experiment was then conducted and at the end, the completed questionnaires were collected from the participating users. Returning to the IRS, the questionnaire data were entered into the IRS and an electronic version of the documents in the collection was presented to the IRS. The IRS information gathering process was then run to populate the token index. The same set of information needs, queries, and phrase-terms presented in the questionnaires was then presented to the IRS search engine to populate the query index. Data collection took place by running IRS-H (using the hybrid indexing method) together with the user data. This was repeated for IRS-I (using the inverted indexing method). The final two activities were to generate the performance measurement data for IRS-H versus the user and for IRS-I versus the user. The generated data were now ready for data analysis.

### **3.3 Research philosophy**

According to Saunders et al. (2019), research philosophy is a flexible term used by a researcher in a particular field to describe the development and nature of knowledge in that field. The development of a research philosophy encompasses the ontological<sup>51</sup>, epistemological<sup>52</sup>, and axiological<sup>53</sup> approaches. Saunders et al. (2019:130) have created a research onion (Figure 3.5) where the ideas and concepts of research philosophy, approach, methodological choice, strategies, and techniques are positioned in a simple way.

---

<sup>51</sup> The branch of metaphysics dealing with the nature of being

<sup>52</sup> The theory of knowledge regarding methods, validity, scope, and distinction between belief and opinion

<sup>53</sup> Relating to the study of values

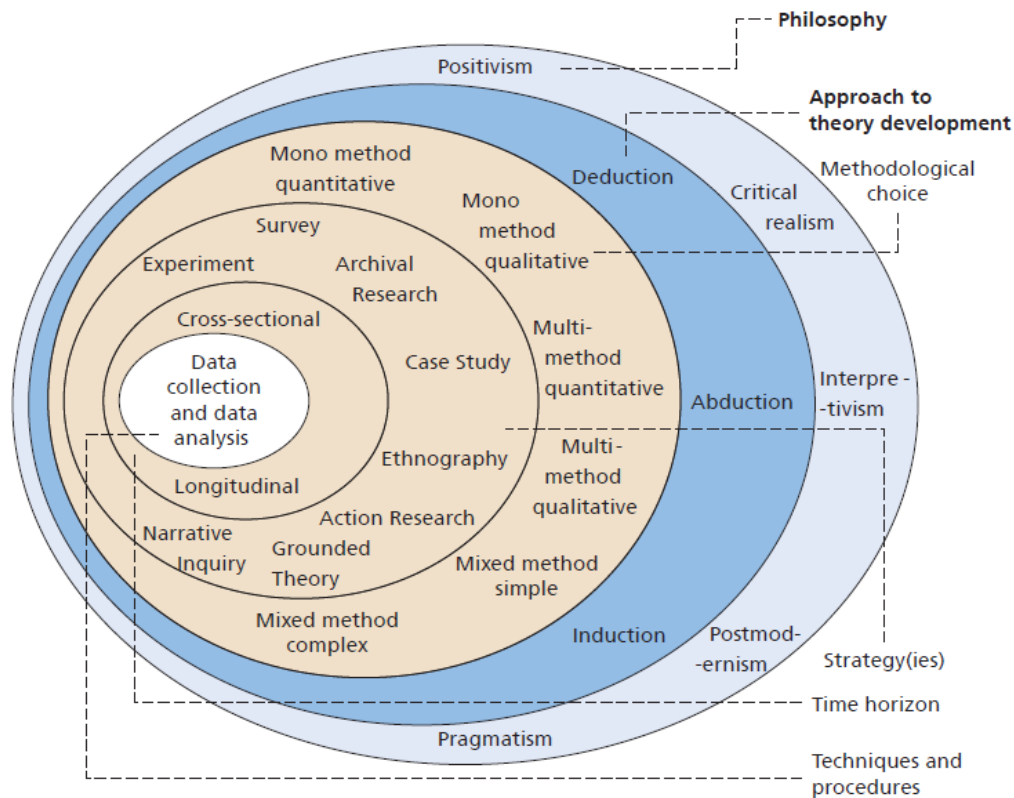


Figure 3.5: The research onion (Saunders et al., 2019:130)

### 3.3.1 Ontology

The branch of philosophy referred to as ontology deals with what objects are and how they exist. One can talk about how the world is, and one can put objects into ontological categories where they seem to belong to each other, all having something in common, so a reality can be shared with others in an effective way. Popper (1978) and Mouton (2004) describe ontology as three worlds (Figure 3.6).

World-1 is where the world consists of physical objects and events; these are objects that people call brute facts (Popper, 1978; Gregor, 2014). According to Mouton (2004), World-1 encompasses the social and physical reality in which individuals exist. Everyday life that produces knowledge of various kinds is often referred to as lay knowledge. This lay knowledge of wisdom, insight, and self-knowledge is gained through experience, learning, and self-reflection (Mouton, 2004).

World-2 is another world about mental objects and events. These objects and events go on inside one's head so these are not observable by anyone else unless they are shared in some way (Popper, 1978; Gregor, 2014). World-2 is the world of scientific research, knowledge and disciplines whereby the researcher selects a phenomenon from World-1 and makes this a subject of organised inquiry, enabling meticulous enquiry into the truth. This truth is referred to as epistemology or *truthful knowledge*

from the Greek word *episteme* (Mouton, 2004). It is a methodological world where the research design is expressed.

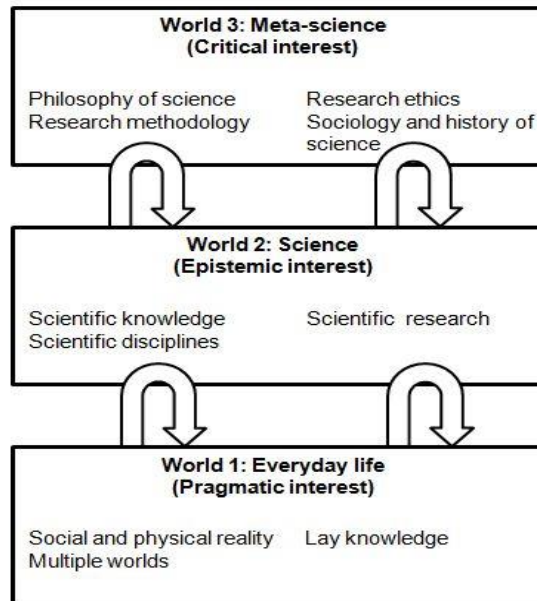


Figure 3.6: A three-world ontology (Mouton, 2004:139)

World-3 is the world of abstract objects such as theories, social institutions, ethics, mathematics, languages, and poems. This is objective-subjective as these objects are what people have invented, for example, mathematics and language; it is hard to get by in life without recognising and sharing these objects. We cannot talk, communicate, or write if we humans do not have words. All these objects in World-3 are constructed socially or individually in a way that help us make sense of the world, and it is imperative to categorise these objects so we can study them in different ways (Popper, 1978; Gregor, 2014). According to Mouton (2004), World-3 is referred to as *meta-science* where past research can be evaluated, and arguments can be made in a way that promises to reveal new insight and improve knowledge of the truth (Mouton, 2004).

From a social science perspective Burrell and Morgan (1979) describe ontology as four research domains (paradigms) derived from consideration of two 2x2 dimensions seen as a 2x2 matrix. These four domains, illustrated in Figure 3.7, consist of the radical humanist, the radical structuralist, the interpretative paradigm, and the functionalist paradigm, and the 2x2 dimensions are radical change and regulation, and objectivism and subjectivism. According to Saunders et al. (2009), these paradigms assist researchers in clarifying their assumptions about reality in social science, function as a tool to facilitate understanding of the work of other researchers, and to assist researchers with navigating their research journey.

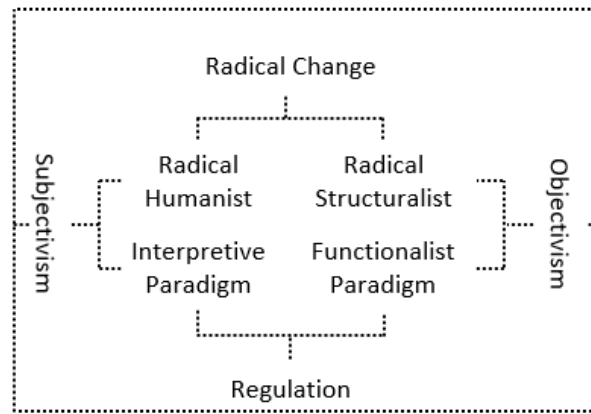


Figure 3.7: Four paradigms for the analysis of social theory (Burrell & Morgan, 1979:23)

Burrell and Morgan (1979) argue that the functionalist paradigm in its most fully developed form generates regulatory sociology and the rebuff is radical change using the theory of conflict rather than the theory of order. Located between the dimensions of subjectivism and regulatory sociology, the interpretive paradigm is an understanding of how the world is at a level of subjective experience and the radical humanist paradigm attempts to develop a sociology between the dimensions of subjectivism and radical (Burrell & Morgan, 1979). Saunders et al. (2019), describe objectivism as the existence of social entities in a reality that is both external to and independent of social actors who are concerned with their existence. It is not subjective when people attach understandings and meanings to social phenomena. The phenomena are created from opinion and the interpretive paradigm (processes) and are constantly reviewed as new information and knowledge becomes available (Saunders, et al., 2009).

The ontological approach for this study therefore falls within the ambit of the objectivist-regulatory-functionalist paradigm: objectivist as this researcher invents a novel design in an attempt to overcome ineffective current indexing methods. Therefore, this researcher designs and builds novel artefacts that provide a successful solution to a known problem; from the design and build of these artefacts, regulatory, so as to control a specific method of doing something; and functionalist, the view of the mind, saying what currently exists is not good enough and must be improved upon through the design of new novel artefacts.

### 3.3.2 Epistemology

According to Saunders, et al. (2009) the branch of philosophy referred to as epistemology is about knowledge and how things are done. Burrell and Morgan (1979) describe epistemology as assumptions about the grounding of knowledge,



understanding the world, the communication to people of this knowledge, and 'what is true' and 'what is false' can be clearly defined. To understand what epistemology is, Saunders et al. (2009) suggest the following question should be asked: what constitutes acceptable knowledge in a field of study? The five epistemological approaches to be considered are:

### **3.3.2.1 Positivism**

Burrell and Morgan (1979) summarise positivism as an attempt to apply methods and models derived from the natural sciences and to treat the social world as if it were a natural world utilising a realistic approach. Saunders et al. (2019) echo these ideas and suggest that a researcher who adopts positivism will usually work traditionally as a natural scientist. Mouton (2004) and Myers (2004) argue that positivism is perceptual experience where reality is objectively given and the measurement instruments based on indefinite variables are independent of the researcher usually using the quantitative method. In positivism, the strategy used will most likely employ an existing theory to develop hypotheses, which will be tested developing further theory that can be tested in the future and that the end product of such research could be regulatory generalisations. The positivist researcher is likely to use a very structured methodology to enable the replication of their work, make quantifiable observations, and perform statistical analysis on the data collected.

### **3.3.2.2 Post-positivism**

According to Creswell (2013), post-positivism represents the thinking after positivism by challenging the out-dated view of the absolute truth of knowledge. However, Ryan (2006a, 2006b) states that the principles of post-positivist research call attention to the creation of new knowledge and meaning. These principles support movements that seek to change the world. Two important characteristics of post-positivist research are firstly, the research is wide ranging as opposed to discipline-based, and secondly, the theoretical realm and the observed realm are intimately interdependent and cannot be isolated. Post-positivism is about understanding how reality is constructed in the knowledge that research is influenced by the researcher's values and theoretical frameworks used (Ryan, 2006a, 2006b; Creswell, 2013).

### **3.3.2.3 Interpretivism**

Interpretivism is about understanding the differences between people as social actors. Critically, the interpretivist researcher must assume an adopted position so as to perform research from their point of view. Interpretivism, the philosophy of law, focuses on the attempts made to understand the meanings that people assign to phenomena. There are no identifiable variables as the focus is on the complexity of

making sense of the situation through the qualitative paradigm (Mouton, 2004; Myers, 2004; Gregor, 2006; Saunders, et al., 2019).

#### **3.3.2.4 Critical realism**

According to Guba and Lincoln (2005), several versions of critical realism theory exist, which include classical critical theory, post-positivist, post-modernist, post-structuralist, and constructivist. More recently, critical research has become an important aspect of information systems research; it is concerned with social issues, including system development, use of systems, and impact on information systems (Myers & Klein, 2011). Guba and Lincoln (2005) suggest four epistemological approaches for research: critical, positivist, post-positivist, and constructivist. In critical research, social critique is used to highlight dominant, restrictive and alienating conditions (Myers, 2004) and it makes use of the participatory paradigm of the participant (Mouton, 2004; Myers & Klein, 2011). Orlikowski and Baroudi (1991) explain in their work that critical research within the information systems world encompasses the following: the phenomenon of interest is single, tangible and can be subdivided; there is a unique description for any chosen aspect of the phenomenon; the researcher and the object of inquiry are independent; a clear dividing line between observation reports and theory statements exists; the possibility of statements generalising laws independent of time or context (nomothetic statements) exists; scientific concepts can therefore be precise having never changing meanings; and unidirectional cause-effect relationships can exist and may be tested using hypotheses and deductive reasoning.

#### **3.3.2.5 Pragmatism**

Saunders, et al. (2019) suggests that there could be more than one way to state a researcher's position, and that it is acceptable to reason that questions of method are secondary to questions of ontology, epistemology, and axiology. Creswell (2013) agrees and suggests that pragmatism is not committed to any one philosophy and reality. This is the position of the pragmatist, and for a pragmatist, the research question is the important aspect in determining the epistemology, ontology, and axiology to adopt. The research question drives which research philosophy to adopt, and whether it is interpretivism or positivism, it is acceptable to perform research with variations in a researcher's epistemology, ontology, and axiology. Creswell (2013) argues that since researchers have a freedom of choice with pragmatism and because a single method is not always applicable, then, for example, mixed-methods research can be applied, drawing assumptions from both qualitative and quantitative data.

This researcher, from an epistemology perspective, adopts the positivistic approach that utilises a strategy using existing theory to develop and test hypotheses. A structured methodology is used to replicate processes and to perform statistical analysis on the data collected.

### **3.3.3 Axiology**

Heron (1996) and Saunders, et al. (2009) suggest that axiology is a branch of philosophy that explores studying the nature of value, goodness, value judgements, and the kinds of things that are valuable. In research, Saunders, et al. (2009) suggest, axiology focuses on the roles that values play in our research choices, for example, in the fields of ethics and aesthetics, and the credibility of the research results. Heron (1996) suggests that researchers reveal their axiological ability when making judgments on how they will approach the research and how they will do it. Hanid (2014:11) describes the basic belief of axiology as, “what is value”. In conclusion of their work entitled, *‘Establishing reliability in design science research’*, and referring to their assessment approaches, Baskerville, Kaul and Storey (2017) suggest that in future research it would be very useful to examine a concept such as axiology. Vaishnavi and Kuechler (2004) and Smuts (2011) suggest that the axiology approach for DSR consists of three things: control, creation, and understanding.

In this study, one could ask, what value or goodness does this research provide to the world? This research consisted of three concepts: control, creation, and understanding through the design, build, and evaluation of a new method of hybridised indexing where better value or goodness is released when using this new method.

This researcher, from an axiology perspective, accepts he is value-orientated and that the research will be interpreted from personal life experiences. In addition, this researcher understands and accepts that his value judgments could have an impact on the control, creation, understanding, and conclusions drawn during this research study.

### **3.4 Research design**

The design for this research encompassed seven main design concepts as presented in the simple flow diagram (Figure 3.8). Stating the research problem was the first concept (section 1.3), followed by a comprehensive literature review (Chapter Two).

The design for this research was dual purpose. It performed:

- i) An exploratory study based on design science research (DSR) (section 3.5.1), in order to:
  - design and physically build an IRS,
  - design and build a new indexing method utilising a pair of hybrid indices,
  - review the literature numerous times gaining insight into existing theories, and
  - perform pilot tests using various text based document collections.
- ii) An explanatory study by conducting an experiment (section 3.5.2):
  - where phrase-terms are expressed as queries,
  - the phrase-terms are applied to the IRS search engine,
  - the phrase-terms are applied to the user questionnaire, and
  - both indexing methods are tested, the hybrid (IRS-H) and inverted (IRS-I), and comparisons made between the data generated by these methods.

Data collection (section 3.6) was thus multi-method quantitative to accommodate both research strategies of an exploratory study using DSR (IRS) and an explanatory study using experimentation (IRS and questionnaire).

The outcome from DSR was to answer the research question relating to IRS design, and the outcome from experimentation using quantitative data analysis (section 3.7) is twofold:

- i) to compare IRS-H with IRS-H and test three hypotheses, and
- ii) to compare IRS-H with the user and test two hypotheses.

This study necessitated an integration of research design choices sourced from the literature. Those used in this study are illustrated in Figure 3.9.

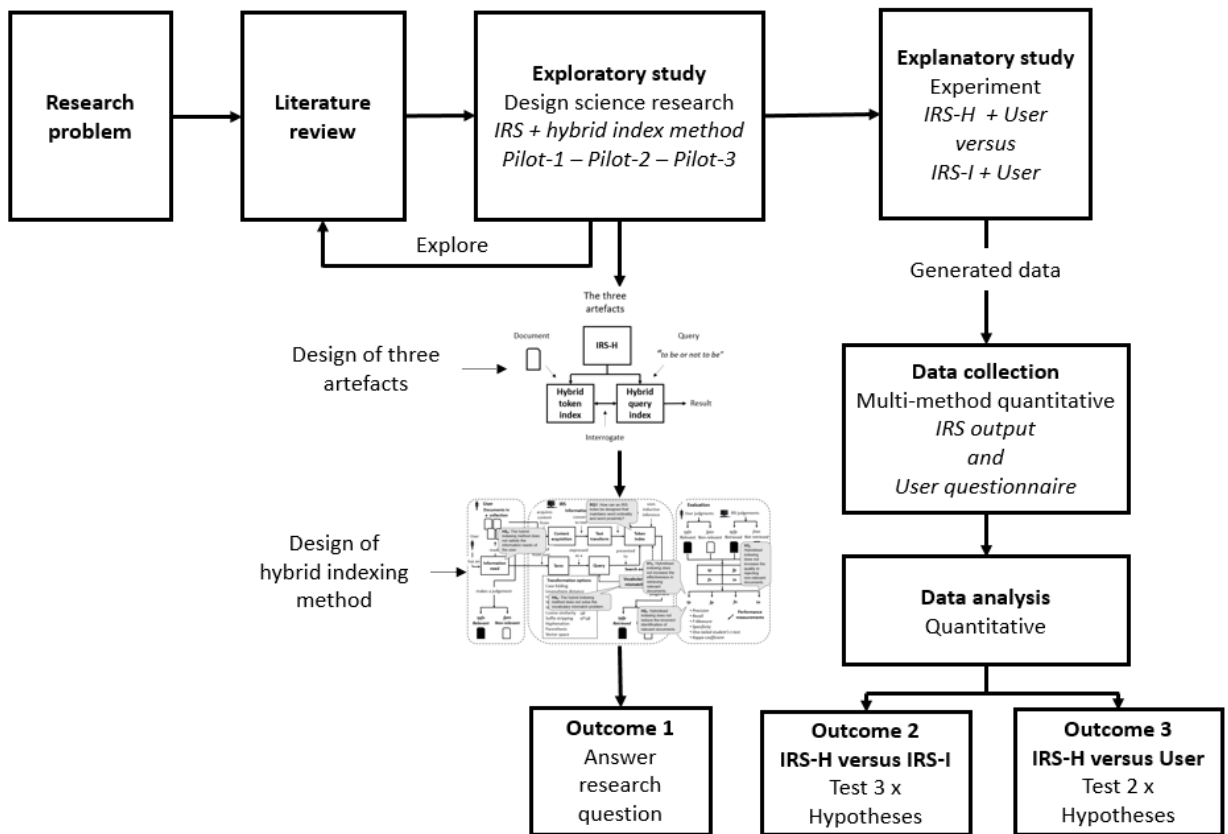


Figure 3.8: A simple flow diagram for this research

The research purpose was both an exploratory study (to design and build an IRS) and an explanatory study to analyse the data generated (section 3.2). The research strategy was DSR (to design and build an IRS) (section 3.5.1) and by experiment to generate data from the IRS and a user completed questionnaire (section 3.5.2). The research approach used deductive reasoning (section 3.5.3) while the research choice was multi-method quantitative (section 3.5.5). The research method was quantitative, utilising data generated by the IRS and questionnaire (section 3.5.4). Finally, the time horizon was cross-sectional as the data generated was a snapshot in time (section 3.5.6).

In order to deal with these design choices, the mapping of these design choices to needs, together with explanatory notes, is presented in Table 3.1.

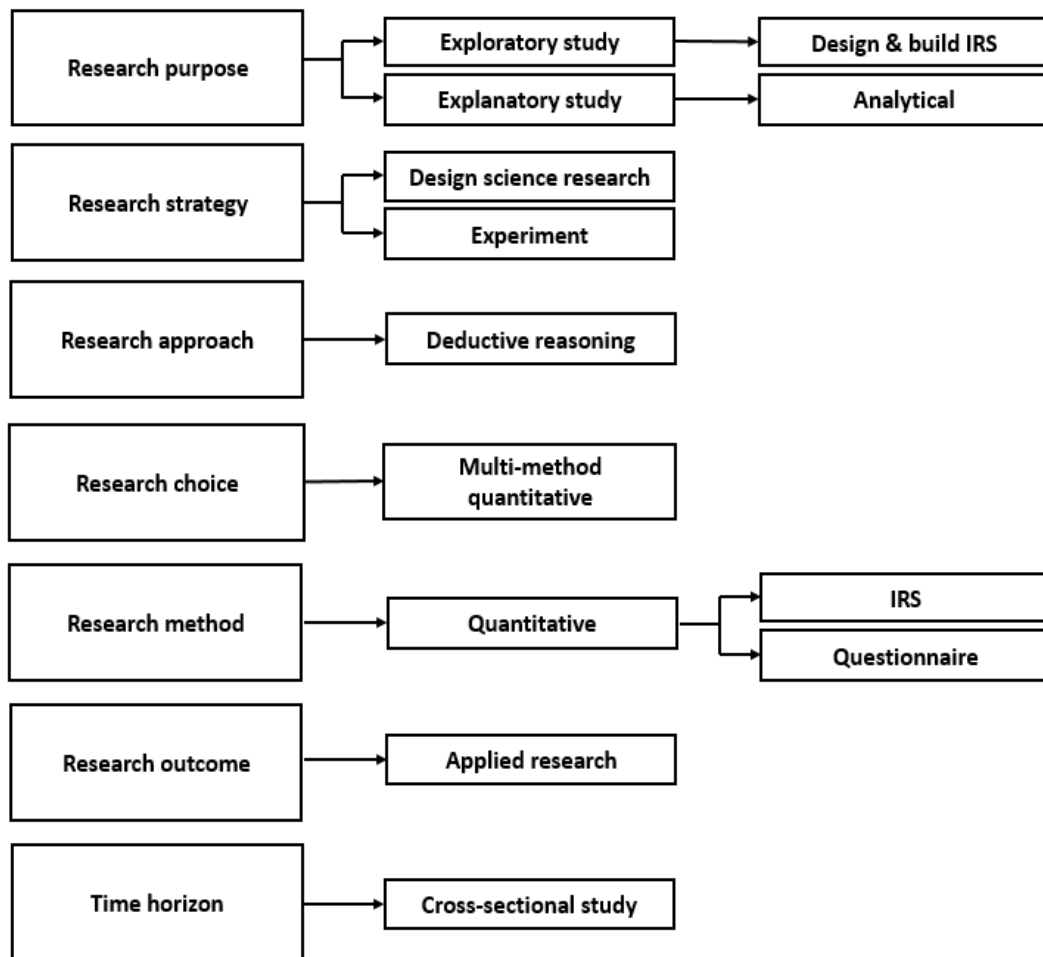


Figure 3.9: A graphic illustration of the research design for this study

Table 3.1: The design choices

Design choice	IRS	User
Exploratory	From the literature, there was a need to determine what IRS theories, concepts, and indexing methods existed and how they could be utilised in this research. In addition, to determine how IRSs make judgements based on these indices. Refer section 3.2.1.	People are never certain and the design options need to be explored. From the literature (Croft, 2019; Vaishnavi, Kuechler, & Petter, 2019) there was a need to determine what role the user played in IRS evaluation and how users make their judgements. Refer section 3.2.1.
Explanatory	Via experimentation, causal relationships between variables needed to be explained to prove a method worked.	Via experimentation, causal relationships between variables needed to be explained to prove a method worked.
Positivism	Reality is objectively given.	Reality is objectively given.
Deduction	A theoretical method where hypotheses were stated to reach a logical conclusion.	A theoretical method where hypotheses were stated to reach a logical conclusion.
Multi-method quantitative	Quantitative analysis of effectiveness was required. Quantitative method using two data collection techniques (from two systems): i) binary data and ii) one set of analysis procedures.	User input to determine effectiveness was required. Quantitative method using one data collection technique (from five users): i) binary data and ii) one set of analysis procedures.

Design choice	IRS	User
Design science research	This research strategy is appropriate for systems design as it uses the cyclical and iterative processes required for the design, build, and test phases of this study. This study falls within the ambit of 'high application domain maturity' and 'low solution maturity'.	N/A
Experiment	A new IRS design (IRS-H) was developed and then tested against an existing design (IRS-I). Therefore IRS-H and IRS-I were fully tested by generating their own system binary data for analysis. Refer section 3.5.2.	Users deliver binary and other numerical data for analysis and therefore a one-day experiment was held to collect user judgments from five participants through the use of a predefined questionnaire. Refer section 3.5.2.
Cross-sectional	System-generated data – a snapshot in time.	User-generated data – a snapshot in time.

### 3.5 Research strategy

The literature revealed that the two most appropriate research strategies in the present context were action research and DSR. Action research involves a cyclical process and so the results from a first cycle (or pilot test) can be used to test a second cycle (a second pilot test), and then repeated thereafter. However, in the progression of this study there were issues, for example, where a design component was found to be missing and as this was an exploratory and explanatory research project, a new artefact had to be created which did not currently exist (sections 4.2 and 4.3). Therefore, DSR was explored more deeply because, in design science, attempts are made to build and develop artefacts empirically that serve human purposes (March & Smith, 1995; Baskerville et al., 2017; Elragal & Haddara, 2019; Thuan, Drechsler & Antunes, 2019) and design science is a problem solver as it creates innovative artefacts to solve real world problems (Hevner et al., 2004; Gregor & Hevner, 2016; Hevner, Vom Brocke & Maedche, 2019).

In the next sections, the following are discussed: i) the design science research strategy; ii) the experiment; iii) the research approach; iv) the research method; v) the research choice; and vi) the research time horizon.

#### 3.5.1 Design science research strategy

The research strategy, chosen for the exploratory purpose of this research, was DSR (Gregor, 2006; Gregor & Jones, 2007; Gregor & Hevner, 2013a, 2013b; Hevner, 2015a; 2015b; Gregor & Hevner, 2016; Gregor et al., 2016; Hevner et al., 2019). DSR can be considered as a set of synthetic and analytical techniques and perspectives that complement the interpretive and positivist perspectives, where phenomena are

created rather than naturally occurring (Vaishnavi & Kuechler, 2004). These phenomena can be sets of interesting behaviours to the researcher, expressed as true knowledge that contribute to knowledge, typically in the form of theories. Design refers to the creation of something new, an artefact, which does not currently exist (Vaishnavi & Kuechler, 2004). Design research is not limited to engineering, science, and information retrieval; it has a wide application (Baskerville et al., 2017; Hevner et al., 2019; Thuan et al., 2019).

The three cycles of DSR are illustrated in Figure 3.10. Hevner (2007) explains that in DSR, the researcher must be active in all three of these cycles during the research project. The first is the relevance cycle that bridges the contextual environment of the research with DSR activities; the second is the rigour cycle that connects the DSR activities with the knowledge base – the scientific foundations of theories, methods, experience, and expertise. These all inform the research project. The third is the design cycle, in the centre, refining core activities of building and evaluating (testing) the design artefacts together with the processes of the research. This is where the challenging part of the research is performed. Hevner (2007) emphasises that in a DSR project, these three cycles must be present and clearly identified.

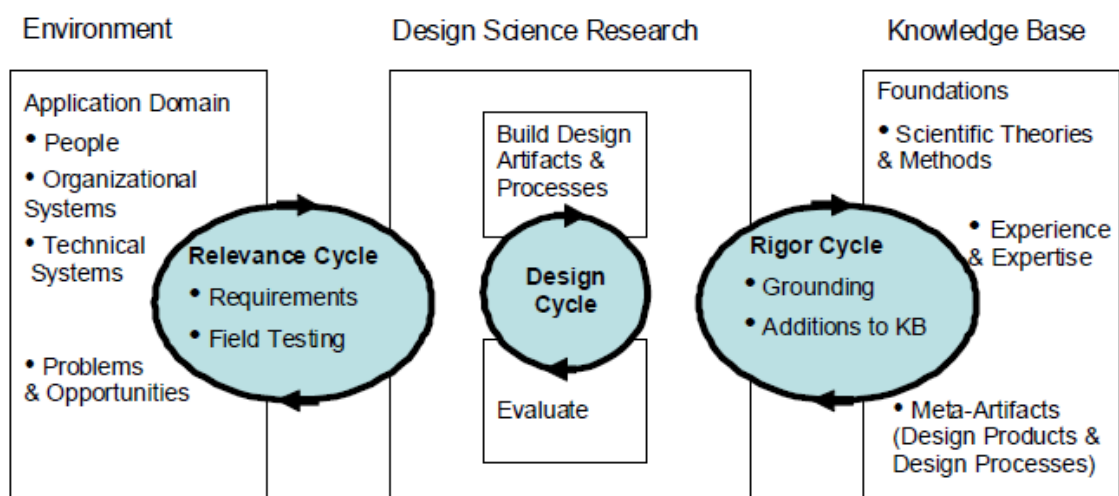


Figure 3.10: Design science research cycles (Hevner, 2007:2)

Based on the original reasoning on design cycle model of Takeda et al. (1990), and using adaptations from the design science research process model of Hevner et al. (2004) and Hevner (2007), and the concept of circumscription by McCarthy (1980), the five process steps followed are:

- i) Awareness of the problem – this is the first process for the research model. DSR is occasionally referred to as *improvement research* and therefore



includes some form of problem solving or improvement method to solve the problem (Vaishnavi & Kuechler, 2004). When information retrieval effectiveness can be improved, there is a research opportunity to improve the effectiveness of an IRS.

- ii) Suggestion – this is the second process for the research model that follows the awareness of the problem. Suggestion is a creative step that provides some insight into a novel design of artefacts for new and existing elements early in the research process. In this research, the aim is to improve efficiency suggesting the design and build of three artefacts: the two hybrid indices and the IRS.
- iii) Development – adaptations to the pilot design will necessitate further developments of the artefacts. During these developments, the design often needs to have its correctness proved by testing the functioning artefacts. The three artefacts will follow a development cycle until functionally acceptable – where the words within the phrase query retrieve documents that contain all the query’s words and in the correct word order (Croft et al., 2015).
- iv) Evaluation – once constructed, the artefacts must be evaluated through testing and the performance, efficiency, and preciseness of the artefact must be analysed by using test collections. The primary test collection consists of 75 queries containing multiple phrase-terms. If additional challenges emerge through circumscription, an iterative process can take place where suggestions to overcome these challengers, described as new knowledge, are made, thus repeating the suggestion-development-evaluation processes. Performance measurements for evaluation include Precision, Recall, F-measure, and Specificity (refer to section 3.7.1 for full measurement details and formulae).
- v) Conclusion – This is traditionally the final process of the research process where conclusions are drawn based upon the findings of the research and recommendations provided.

### **3.5.1.1 The three design cycles**

The three design cycles for this study are illustrated in Figure 3.11. The black images represent the number of documents in each collection while the white images represent the processes.

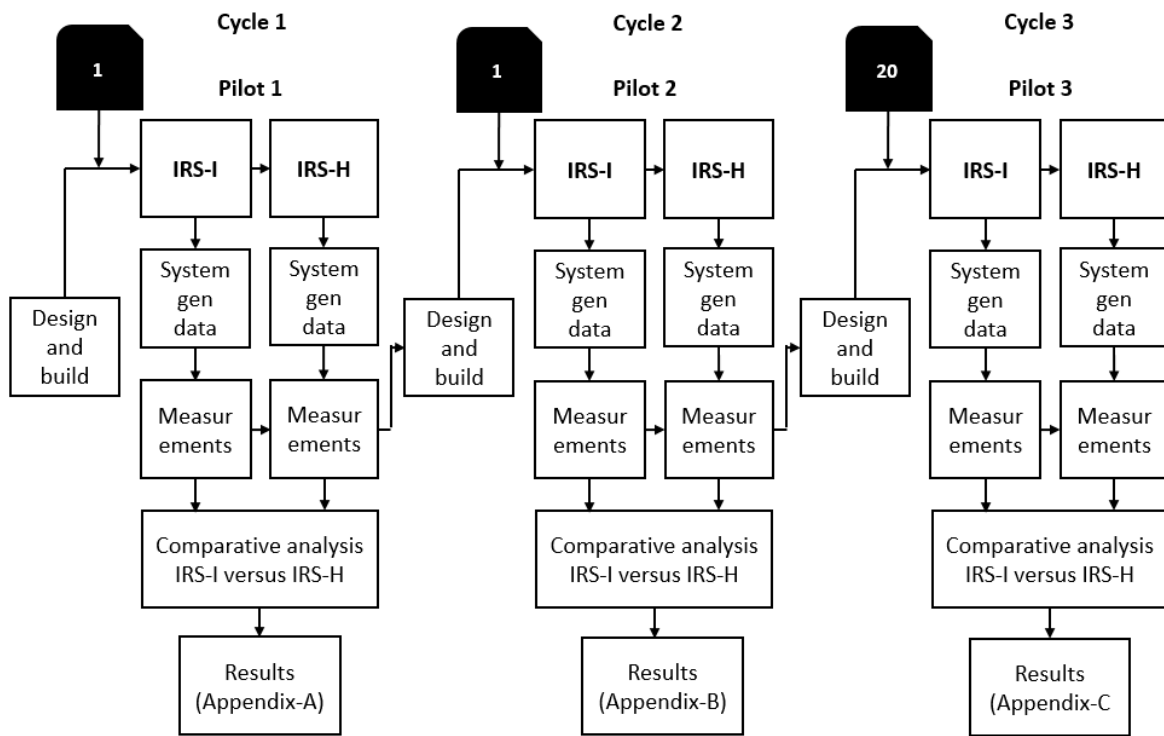


Figure 3.11: The three design cycles for this study

The first cycle represents Pilot 1 where the original design and build took place. One IRS was built using the inverted indexing method, and is referred to as IRS-I. The second was cloned from IRS-I, and the design was adjusted to accommodate the hybrid indexing method, and is referred to as IRS-H. Both these IRSs in Pilot 1 used the same single document for testing. Both IRSs generated data and computed performance measurements. The main usage difference between the two IRSs is that IRS-H used multi-word phrase-terms rather than single-word terms in its queries to represent an information need. IRS-H therefore used phrase-term frequency (*ptf*) rather than term frequency (*tf*). All other system-generated data concepts were identical between the two IRSs.

The performance measurements used to evaluate the IRSs were also identical. From these measurements, any design issues were documented and where necessary, the IRSs were rigorously redesigned and rebuilt. Cycle 1 was then repeated as Cycle 2 (referred to as Pilot 2) with the design changes in place and both IRSs were tested using a single document. Both IRSs generated data and computed performance measurements. Further design changes were made and Cycle 2 was repeated as Cycle 3 (referred to as Pilot 3). With updated design changes, both IRSs were again tested but this time they both used a collection of 20 documents. Both IRSs generated data and computed performance measurements for this last testing cycle.

### 3.5.1.2 Test collection preparation

For this study, utilising the DSR cycles of design, build and evaluation, three test collections were used to test the two IRSs rigorously: IRS-I used the inverted index method and IRS-H used the hybrid index method.

The first test collection that was pilot tested, referred to as Pilot 1, comprised two pages from one document, Hamlet Act 3 Scene 1 written by William Shakespeare (2018). This book was specifically selected for its Elizabethan English / Early modern English and catchy phrases. Four information needs were applied, expressed as four queries. During hybrid index evaluation, these queries used three multi-word phrase-terms and during inverted index evaluation, eight single-word terms were used. The design, build, and evaluation of Pilot 1 are discussed in Appendix A.

The second test collection, referred to as Pilot 2, comprised one 666-page document, the book '*Ulysses*', written by James Joyce (1932). This book was selected for the author's use of unimaginable phrases, length of words, morphemes<sup>54</sup>, and phonemes<sup>55</sup>. In total, 26 information needs were expressed as 26 queries. During hybrid index evaluation, these queries used 26 phrase-terms, six of which were lengthy single-word phrase-terms, and during inverted index evaluation, the queries used 46 single-word terms. The design, build, and evaluation of Pilot 2 are discussed in Appendix B.

The third test collection referred to as Pilot 3, explored the problem of vocabulary mismatch, and comprised 20 documents within the collection. In total, 14 information needs were expressed as 14 queries. During hybrid index evaluation, these queries used 14 phrase-terms and during inverted index evaluation, the queries used 15 single-word terms. The design, build, and evaluation of Pilot 3 are discussed in Appendix C.

### 3.5.2 The experiment

The research strategy, chosen for the explanatory stage of this research, was an experiment (Baskerville et al., 2017) where theoretical hypotheses were presented. A sample of documents (a test collection) from a known population was selected. Two systems IRS-H and IRS-I produced system-generated outputs and a group of participants completing questionnaires produced the user output. The experiment is

---

<sup>54</sup> A meaningful morphological (the study of words) unit of a language that cannot be further divided

<sup>55</sup> A distinct unit of sound in a specified language distinguishing one word from another

discussed in detail in section 3.2.3 and the data generated by the experiment in section 3.6.

### **3.5.2.1 Test collection preparation**

The test collection used in the experiment during system evaluation comprised 100 systematically, randomly sampled documents. In total, 75 information needs were expressed as 75 queries. During hybrid index evaluation, these queries used 65 phrase-terms and during inverted index evaluation, the queries used 49 terms.

### **3.5.2.2 The questionnaire**

This section describes how the user judgement experiment unfolded. Prior to the experiment, this researcher and a nominated facilitator prepared 100 printed documents and marked these documents with their corresponding unique document numbers. In addition, five sets of questionnaires were printed, one set per user, pertaining to 75 queries, ten of which represented ten expanded queries and their associated phrase-terms. A large conference room was booked and five users, all with research experience and affiliated with CPUT in some way, were selected to participate in the one-day user judgement experiment. Using stratified sampling, the 100 printed documents were arranged in five piles on a table according to document thickness with the five sets of questionnaires, for selection by each of the five users.

At the beginning of the experiment, methods and techniques were discussed and a brief overview of the experiment, explaining the process, was provided. Demographic data relating to each user were captured, and each participant signed a document, giving their permission for the use of their data (Appendix J). For data analysis purposes, each of the five users was allocated a code from A through to E.

Each of the participating five users randomly selected 20 documents from each of the five piles and was issued a questionnaire (refer to Appendix E for the full questionnaire). Within each section of the questionnaire (each section referring to one information need based on expanded queries ranging between four and ten phrase-terms), the user was required to indicate with a tick or cross whether the document was relevant to the information need, and whether each phrase-term existed, or did not exist, within each of the documents. An example is presented in Figure 3.12.

User -									
In01: I want to find all documents relevant to design science research									
For each of the documents handed out to you please write down the document number in column 1 and thereafter indicate with a tick (true) or cross (false) whether each phrase term pt01 through to pt08 (columns 2 to 9) exists within each of the documents. In addition, in the last column, please indicate with a tick (true) or cross (false) whether each document is relevant to the information need stated above.									
Doc	pt01 design science	pt02 design sciences	pt03 design science research	pt04 design science methodology	pt05 the design method	pt06 design research	pt07 design science research paradigm	pt08 design science research paradigms	Document relevant to information need In01?
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									
d									

Figure 3.12: An example of the questionnaire

### 3.5.2.3 The systems: IRS-H versus IRS-I

To test the two systems for each of the three hypotheses, the control group was IRS-I, and the test group IRS-H. The independent variable in all three tests was hybridised indexing and the three dependent variables were: i) retrieval effectiveness; ii) incorrect identification of relevant documents; and iii) quality in rejecting non-relevant documents (Table 3.2).

Table 3.2: Control and test groups: H1, H2 and H3

Hypothesis	Control group	Test group	Independent variable	Dependent variable
H1	IRS-I	IRS-H	Hybridised indexing	Retrieval effectiveness
H2	IRS-I	IRS-H	Hybridised indexing	Incorrect identification of relevant documents
H3	IRS-I	IRS-H	Hybridised indexing	Quality in rejecting non-relevant documents

Figure 3.13 presents how system-generated data during the experiment was prepared and how the experiment was conducted. During the experiment, data were generated by running IRS-H together with the user data to produce specific key values. This generation of data was repeated for IRS-I with the user data thus

producing two system sets of data allowing IRS-H to be compared with IRS-I. The key values were  $tp$ ,  $tn$ ,  $fp$ ,  $fn$  and  $ptf$ , which could now be applied to the various formulae that produce the performance measurements for Precision, Recall, Specificity, the phrase-term-by-document matrix, and  $tp$ ,  $tn$ ,  $fp$ ,  $fn$  and  $ptf$ . These performance measurements were then used to test the first three hypotheses of this study: **H1**, **H2**, and **H3**.

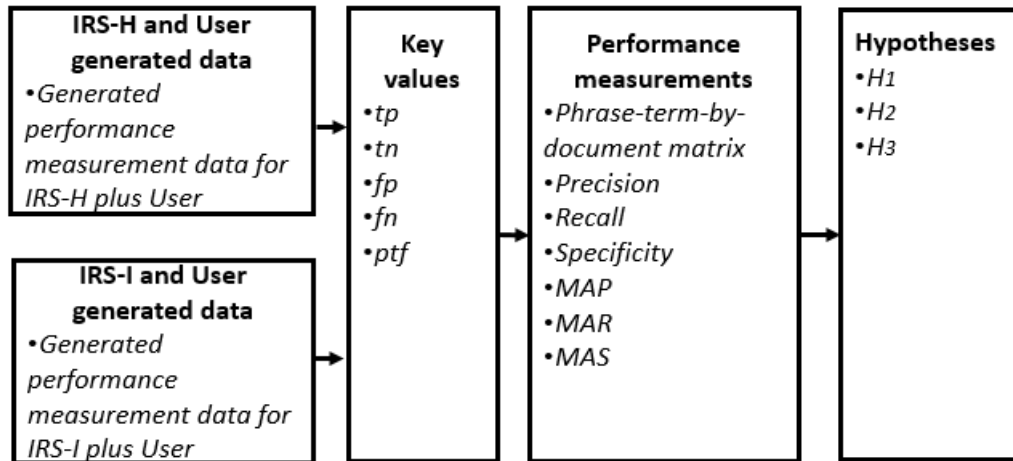


Figure 3.13: A flow chart representing system-generated data

#### 3.5.2.4 The systems: IRS-H versus user

To test the two systems for each of the three hypotheses, the control group was the user and the test group was IRS-H. The independent variable in both tests was the hybrid indexing method and the two dependent variables were: i) agreement in judgements; and ii) satisfying the information needs of the user (Table 3.3).

Table 3.3: Control and test groups: H4 and H5

Hypothesis	Control group	Test group	Independent variable	Dependent variable
H4	User	IRS-H	The hybrid indexing method	Agreement in judgements
H5	User	IRS-H	The hybrid indexing method	Satisfying the information needs of the user

Figure 3.14 presents how system-generated judgements and user-generated judgements data during the experiment were prepared, and how these judgements were analysed. During the experiment, data were generated by IRS-H to produce judgement values ( $tpfp$  and  $fnfn$ ) and by analysing the questionnaire to produce user judgement values ( $tpfn$  and  $fpfn$ ). The judgment values were applied to the group formulae that produce the judgment measurements using the Kappa coefficient and

an agreement measurement scale. These judgement measurements were then used to test the final two hypotheses of this study: **H4** and **H5**.

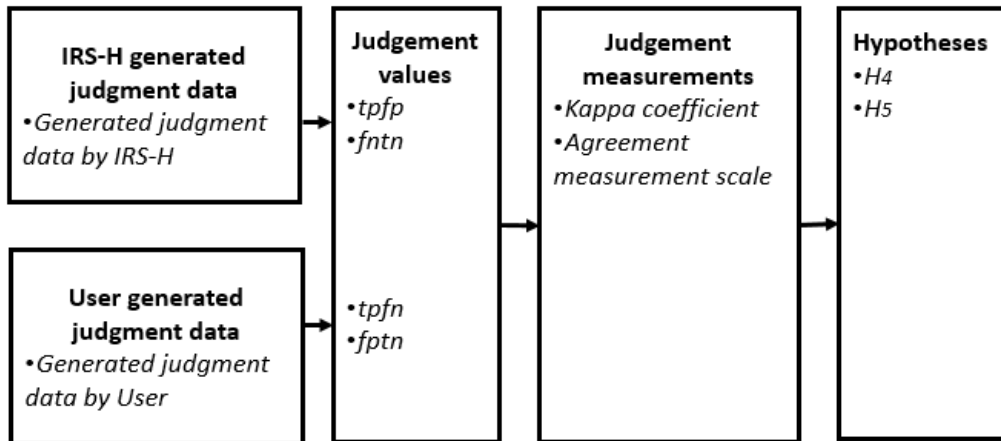


Figure 3.14: A flow chart representing system and user-generated judgments

### 3.5.3 Research approach

This study took place in a context of ‘improvement’ where a new solution was developed for a known problem. Within the knowledge contribution framework, this study falls within the ambit of high application domain maturity and low solution maturity, thus providing both a research opportunity and a knowledge contribution (Hevner et al., 2004; Hevner, 2007).

This study made use of deductive reasoning using a top-down “theory → hypothesis → observation → empirical generalisations” approach (Babbie, 2013:22) to reach a logical true conclusion (Pierce, 1958:46). The study was based on a clear existing theoretical position where hypotheses were stated, variables measured, and outcomes examined, and where the theoretical laws presented the basis of explanation (Saunders et al., 2019).

### 3.5.4 Research method

The research method used in this study was a quantitative research method using numerical methods and statistics.

### 3.5.5 Research choice

The multi-method quantitative choice was selected for this research study, where three data collection techniques and one set of analysis procedures (quantitative) were used (Gacenga et al., 2012).

### **3.5.6 Research time horizon**

The research time horizon for this research was a cross-sectional study, as the data collected were a snapshot in time.

## **3.6 Data collection**

Saunders et al. (2009) suggest that with positivism, data collection should be highly structured using large samples of quantitative data with specific measurements. The research choice was multi-method quantitative where three data collection techniques and one set of analysis procedures (quantitative) were used.

### **3.6.1 Units of analysis and observation**

King, Keohane and Verba (1994) argue that the differentiating factors between a unit of analysis (UoA) and a unit of observation (UoO) is at data detail level. Using a top-down approach with level 1 at the top, the unit of observation would be below the unit of analysis. Therefore, the unit of observation should exist during data collection as this process takes place before data analysis, and the unit of analysis (Marais, 2016; Bonello & Meehan, 2019) should exist during the data analysis process. Sedgwick (2014) concurs that the unit of observation and the unit of analysis are often misunderstood and clarifies the differentiating factors of the two measurements. According to Sedgwick (2014), the unit of observation statistically speaking defines the 'who' or 'what' when data are collected or measured. However, the unit of analysis statistically speaking defines the 'who' or 'what' when information is analysed and conclusions are made (Sedgwick, 2014). Seddon et al. (1999), referring to the original work of DeLone and McLean (1992) who created six information systems effectiveness categories based on a unit of analysis and evaluation type context dimensions, argue that although there are many measures in the literature, the unit of analysis is seldom found, which hampers the clarity of the research, making the research more difficult for the reader to understand. In the work of Seddon et al. (1999), the authors used a two dimensional matrix that classified information systems success measures where the first dimension represented the stakeholder and the second dimension the system. From their matrix, Seddon et al. (1999) derived 30 measurement classes, each representing a specific unit of analysis. Pather (2006) echoes the work of Seddon et al. (1999) and agrees that the unit of analysis is an important entity in understanding and measuring information systems success. When considering IRSs, Tang (1999) used document evaluation, the dimensions of criteria, and the formats of documents as units of analysis. Recalling the two IRS matrices discussed in Chapter Two (section 2.9), for the first, Kobayashi et al. (2015) described how the term-by-document matrix was used to measure term and document



relationships where the columns represented the terms and the rows represented the documents, and the second, Cleverdon (1967) described the use of the 2x2 contingency table (De Raadt, et al., 2019), where the vertical columns represented the user's judgements and the horizontal rows the IRSs judgements. These matrices were used to understand the performance and effectiveness of IRSs better. In this research, the UoA is a query and the UoO is a document.

### **3.6.2 Sampling techniques**

Trochim (2006) and Kelly (2009) explain that sampling techniques in research can be separated into two types: the first is probability or representative sampling, and the second is non-probability or judgemental sampling, as illustrated in Figure 3.15. Probability sampling (or representative sampling) is associated with experiment and survey research strategies.

An example of non-probability sampling is purposive sampling, also referred to as judgemental sampling, which allows a researcher to use his/her judgement to select cases that help meet the research objectives. Other forms of non-probability sampling are snowball sampling and convenience sampling. Snowball sampling is often used when members of the desired population are difficult to identify. Convenience sampling (or haphazard sampling) involves selecting conveniently positioned cases at random. Simple random sampling is a technique whereby cases are selected using computerised random number generators or simple random number tables. Creswell (2013) suggests that instead of tables or number generators, the researcher can systematically choose, at random, the beginning of a list and then select every  $n$ -numbered item on the list, where  $n$  is a fraction determined by the number of cases.

Random sampling is best used when an accurate and easily accessible sampling frame exists, listing the entire population (in this study the document collection), which is preferably stored on a computer. Systematic random sampling is a method where every  $k^{\text{th}}$  unit,  $k$  being the interval size, is selected from the population, so that interval size:  $k = \text{population} / \text{sample size}$  (Trochim, 2006; Kelly, 2009; Saunders et al., 2009).

In this research (Figure 3.15), the process flow of sampling techniques was as follows: this researcher held a document collection containing 2,271 documents collected over the past 13 years. From this document collection and for Pilot 1 of this study, a single document containing two pages was purposively sampled. From the same collection and for Pilot 2 of this study, a single document was purposively sampled containing 666 pages. From the same collection and for Pilot 3 of this study, 20 documents were purposively sampled, ten relevant to a set of information needs and ten non-relevant.

A university researcher purposively sampled 896 documents from his own personal document collection, and of these 896 documents, 100 were sampled systematically and randomly by this researcher. In preparation for the systematic random sampling, each of the 896 documents was allocated a sequentially generated document number. These document numbers were then listed in a spreadsheet without any other information. Systematic random sampling was then performed by selecting every 8<sup>th</sup> document number from the list, to create the final collection of 100 documents for the evaluation.

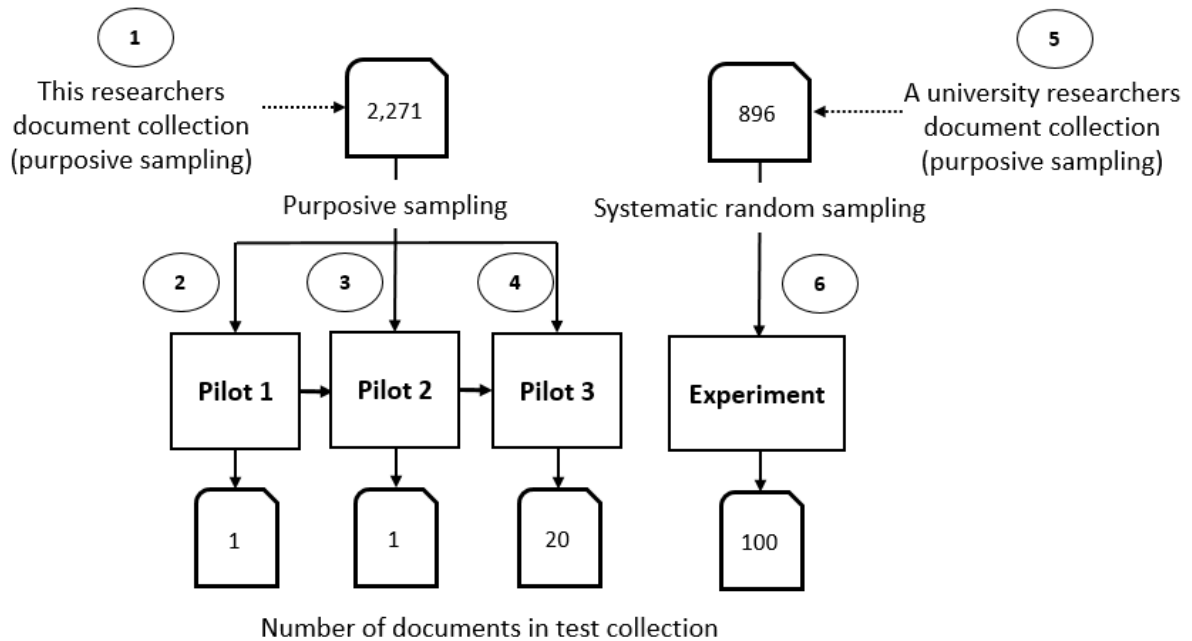


Figure 3.15: Sampling techniques adopted for this research

### 3.6.3 The experiment

For the experiment in this research, the data collection approach was highly structured, as most of the data were system generated by the two IRSs. However, data collection for the user's judgements was collected manually using a predefined questionnaire pertaining to 75 queries (65 single phrase-term queries and ten expanded queries using multiple phrase-terms). System-generated quantitative data were collected from the search engine results produced by the two systems: firstly by IRS-I and secondly by IRS-H. The framework for data collection is illustrated in Figure 3.16 and is followed by various methods and techniques employed in this research.

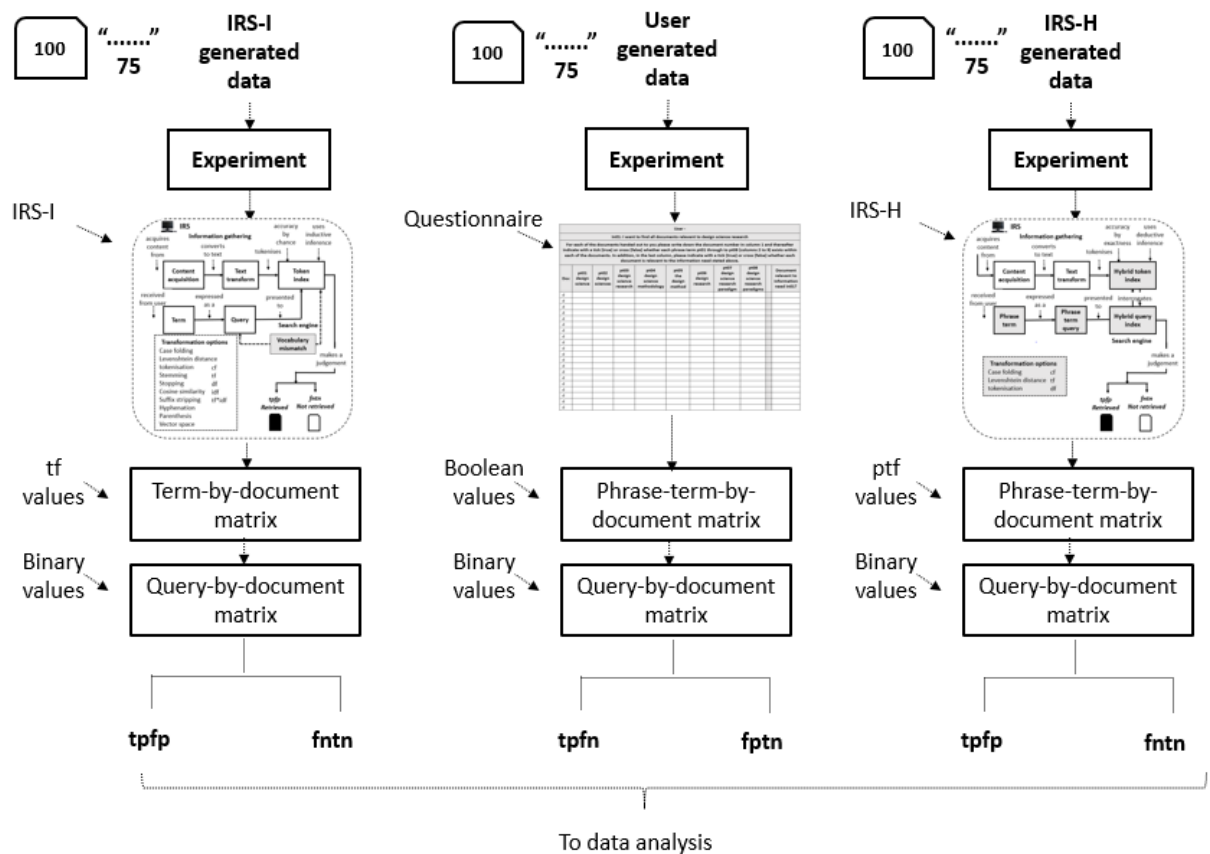


Figure 3.16: The experimental framework

For IRS-I, three rigorous pilots were performed to build and test the system using the inverted index. Once built, the information gathering and the search engine processes were evaluated. These processes generated quantitative data in the form of a term-by-document matrix using term frequency values, which were then converted to binary values and stored in the query-by-document matrix. The generated outputs for data analysis purposes were the values of *tpfp* and *fntn* for each of the 75 queries.

Data collection for IRS-H followed a similar method where three rigorous pilots were performed to build and test the system using the pair of hybrid indices. After the build and successful pilot testing, the information gathering and search engine processes were evaluated. These processes generated quantitative data in the form of a phrase-term-by-document matrix using phrase-term frequency values, which were then converted to binary values and stored in the query-by-document matrix. Again, the generated outputs for data analysis purposes were *tpfp* and *fntn* for each of the queries with values differing from those of IRS-I.

To collect the user-generated data, a ten-page questionnaire (Appendix E) was used to gather user-generated quantitative data during a one-day experiment held at CPUT with five participating users. The Boolean data captured via the questionnaire were

converted to binary values and stored in the query-by-document matrix. The generated outputs for data analysis purposes were the values of *tpfn* and *fptn*.

### 3.7 Data analysis

Extensive data analysis was performed during the three pilot tests during the design, build, and testing cycles. These analyses are presented in Volume II Appendices A, B and C for Pilot 1, 2 and 3 respectively. Supporting the research questions IRS performance measurements were used to judge the effectiveness of the two IRSs. The results from the questionnaire provided user relevant (*tpfn*) and user non-relevant (*fptn*) values in Boolean format, which were then converted to binary where 1 represented *true* and 0 represented *false*. The IRS generated data provided system retrieved (*tpfp*) and system not-retrieved (*fntn*) values provided in term frequency (for IRS-I) and phrase-term frequency (for IRS-H) format, which were then converted to binary where 1 represented  $tf > 0$  (for IRS-I) or  $ptf > 0$  (for IRS-H) and 0 represented  $tf = 0$  (for IRS-I) or  $ptf = 0$  (for IRS-H).

#### 3.7.1 Performance measurements

From the data collection process discussed earlier and with the system and user-generated values of *tpfn*, *fptn*, *tpfp*, and *fntn* now known, the values for *tp*, *fp*, *fn* and *tn* were derived. From the literature, eight performance measurements were used to calculate for each query, the values for: Precision, Recall, Fallout, F-measure, Snobbery ratio, Specificity, Noise factor, and Accuracy (Kent et al., 1955; Cleverdon & Keen, 1966; Salton et al., 1975; Van Rijsbergen, 1979; Kohavi and Provost, 1998; Manning et al., 2008). These measurements for each IRS are presented in table format in Chapter Four and in Appendix I.

#### 3.7.2 Statistical analysis

Statistical analysis<sup>56</sup> was performed to test the three hypotheses. To test the first hypothesis, precision, ranked average precision (AP) and mean average precision (MAP) were utilised (Waitelonis, 2018), and to test statistical significance, a one-tailed

---

<sup>56</sup> Note: IBM SPSS statistics version 25 (SPSS) was used to perform the statistical analyses for the one-tailed student's t-test and the Kappa coefficient. The objective was to perform a one-tailed student's t-test at the 95% confidence level but SPSS could only perform a two-tailed student's t-test. In SPSS, the independent samples t-test was used to check if the two systems (variances) were statistically different from each other. As each system had 75 values, one per query, the sample size (N) was 150 and as there were two systems the degrees of freedom (*df*) equalled the sample size minus the number of systems (on the level of average) in the test, therefore  $df = 150 - 2 = 148$ . To mirror the results on the one-tailed student's t-test in SPSS at a 95% confidence level, the independent samples t-test was performed at the 90% confidence level producing the tails with  $\alpha = 0.05$  rejection region and the significance level *p* was adjusted to half its value. The results produced by SPSS are presented in Appendix K. The statistical analysis for this research was reviewed by a professionally registered statistician at CPUT.

student's t-test was performed – an appropriate statistical significance test for IRSs suggested by Smucker et al. (2007). To test the second hypothesis, Recall, ranked average recall (AR), and mean average recall (MAR) were utilised, and to test statistical significance, a one-tailed student's t-test was performed. To test the third hypotheses, Specificity (S), ranked average specificity (AS) and mean average specificity (MAS) (Choudhary et al., 2017) were utilised, and to test statistical significance, a one-tailed student's t-test was performed. To support the last two hypotheses, the Kappa coefficient (Cohen, 1960; De Raadt et al., 2019) was used to determine any differentiation between user and IRS judgements. Agreement measurements (Smucker et al., 2007; Chaparro et al., 2016) used the six-division range based on the work of Landis and Koch (1977).

The objective of the data analysis in this research was to prove that this novel method worked. In this data analysis section, a number of statistics were produced. These statistics were used to validate what was done, to test the three hypotheses, and to prove that the hybrid indexing method worked. For clarity of purpose, it needs to be said again that these statistics and the conclusions from these statistics are not the contribution to knowledge of this study – the contribution to knowledge here is the hybrid indexing method.

### **3.8 Ethical considerations**

As the researcher for this study, this author acknowledged that it was his responsibility to follow the Cape Peninsula University of Technology code of practice on ethical standards together with any relevant academic or professional guidelines in the conduct of this study. All the computer software, Microsoft Access and Microsoft Visual Basic Access, used in this research was fully licenced. This researcher's number-based ethics lie in the program code developed and how the data have been treated in the development of the three artefacts including the indexing methods. The intellectual property of this research is shared 20% for CPUT and 80% for the author – this was arranged with the CPUT Technology Transfer Office (CPUT, 2019).

### **3.9 Summary**

The chapter began with an overview of the research problem. The research questions and the five hypotheses were presented. This was followed by the research purpose, research philosophy, the research strategy, the data collection techniques, and the method of data analysis. This research was both an exploratory study and an explanatory study based on the quantitative method. The philosophy used the ontological approach of the objectivist-regulatory-functionalist paradigm, as it was an evaluation of the effectiveness of an IRS. The epistemological approach adopted

positivism as the reality was objectively given and the measurement instruments were independent of the researcher and the quantitative paradigm. The research strategy was DSR using the deductive research approach for the development of the artefacts. This research was based on a clear existing theoretical position where hypotheses were stated, the variables measured, the outcome examined, and the theoretical laws presented the basis of explanation. The multi-method quantitative research choice was used for this cross-sectional study. The research method was the quantitative research method using numerical methods and statistics and the research outcome was reached by means of a mixture of both basic research and applied research.

## 4. CHAPTER FOUR: RESEARCH RESULTS

*“To accomplish great things, we must not only act, but also dream; not only plan, but also believe”*

– Anatole France

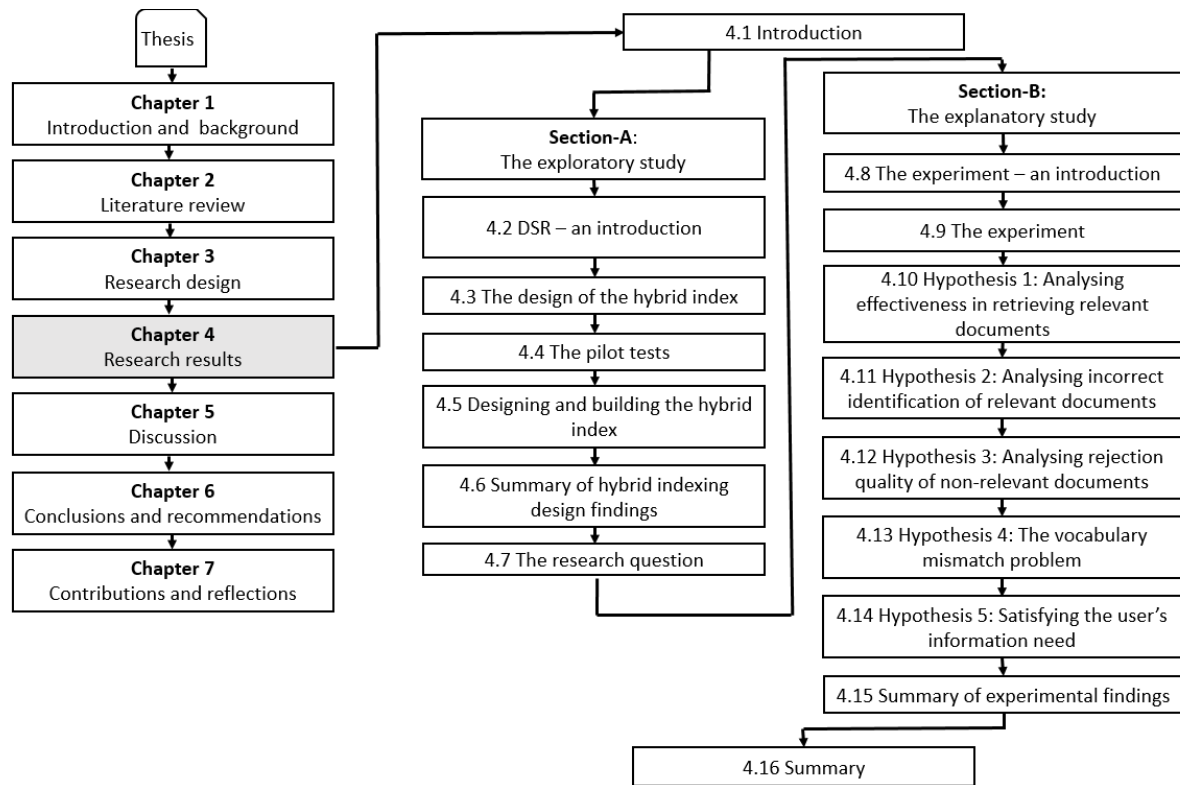


Figure 4.1: Schematic representation of Chapter Four

### 4.1 Introduction

This chapter presents the results from this research and summarises the findings relating to the research questions and the five hypotheses. For the benefit of the reader the research questions, the hypotheses, objectives, and methods together with the sections within this chapter and the appendices to which they pertain (Table 4.1) are restated.

Table 4.1: Research questions, hypotheses, objectives, methods and sections

Research questions / Hypothesis	Aim / Objective	Method	Section
<b>RQ1:</b> How can an IRS index be designed that maintains word ordinality and word proximity?	To design, build, and rigorously pilot test a hybrid indexing method that maintains word ordinality and word proximity, and to compare the effectiveness of this method with the traditional inverted indexing method	Literature review Exploratory Design science research Hybrid index design and build (IRS-H) Perform three pilot tests	Chapter Four Section A sub-sections 4.2, 4.3, 4.4, 4.5, 4.6 & 4.7 and Volume II Appendices A, B & C

Research questions / Hypothesis	Aim / Objective	Method	Section
<b>H1<sub>0</sub></b> : Hybridised indexing does not increase the effectiveness of retrieving relevant documents	To test whether an IRS using a hybrid indexing method increases the effectiveness of retrieving only those documents that are judged relevant by the user	Literature review Explanatory, Experiment IRS-I and IRS-H tests Performance measurements Precision, Ranking, MAP Statistical analysis One-tailed t-test	Chapter Four Section B sub-sections 4.8, 4.9 & 4.10 and Appendices D, E, F, G, H, I & J
<b>H2<sub>0</sub></b> : Hybridised indexing does not reduce the incorrect identification of relevant documents	To test whether the hybrid indexing method reduces errors in incorrect identification of user judged relevant documents, thus reducing the number of documents for the user to peruse	Literature review Explanatory, Experiment IRS-I and IRS-H tests Performance measurements Recall, Ranking, MAR Statistical analysis One-tailed t-test	Chapter Four Section B sub-sections 4.8, 4.9 & 4.11 and Appendices D, E, F, G, H, I & J
<b>H3<sub>0</sub></b> : Hybridised indexing does not increase the quality in rejecting non-relevant documents	To test whether the hybrid indexing method increases the rejection quality of user non-relevant documents, thus providing confidence to the user in the judgement of the IRS	Literature review Explanatory, Experiment IRS-I and IRS-H tests Performance measurements Specificity, Ranking, MAS Statistical analysis One-tailed t-test	Chapter Four Section B sub-sections 4.8, 4.9 & 4.12 and Appendices D, E, F, G, H, I & J
<b>H4<sub>0</sub></b> : Judgments made by the hybrid indexing method and the user disagree	To determine whether the judgments made by the hybrid indexing method and the user agree	Literature review Explanatory, Experiment User judgements IRS-H judgements Kappa coefficient Agreement measurements	Chapter Four Section B sub-sections 4.8, 4.9 & 4.13 and Appendices D, E, F, G, H, I & J
<b>H5<sub>0</sub></b> : The hybrid indexing method does not satisfy the information needs of the user	To determine whether the hybrid indexing method satisfies the information needs of the user by retrieving those documents from the collection that are relevant to the user	Literature review Explanatory, experiment User judgements IRS-H judgements Kappa coefficient Agreement measurements	Chapter Four Section B sub-sections 4.8, 4.9 & 4.14 and Appendices D, E, F, G, H, I & J
<b>RQ2</b> : Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?	To determine whether the hybrid indexing method solves the problem of mismatching vocabulary between a query and a document	Literature review Exploratory and Explanatory results from <b>RQ1</b> and <b>H1</b> , <b>H2</b> , <b>H3</b> , <b>H4</b> and <b>H5</b> and findings	Chapter Four section 4.16, Chapter Five sections 5.2, 5.3, 5.4, 5.5, 5.6 and 5.7

This chapter is presented as two studies, Section A and Section B, with various sub-sections within each of them:

- Section A represents the exploratory study (section 3.5.1) in four sub-sections that used DSR to design and build the IRS together with its hybrid indexing method. The results are used to answer the first research question.



- Section B represents the explanatory study (section 3.5.2) in eight sub-sections that used experimentation to generate and then collect and analyse the data to test the five hypotheses quantitatively. The cumulative results are used to answer the second research question.

## **SECTION A – THE EXPLORATORY STUDY**

### **4.2 DSR – an introduction**

The exploratory study encompassed the designing and building of three artefacts: an IRS, referred to as IRS-H, and a pair of hybrid indices: the first described as the hybrid token index and the second as the hybrid query index, referred to collectively as the hybrid indices.

DSR was used as a research method for the indexing design, which necessitated numerous design cycles to be followed, referred to as pilot tests in this study (section 3.5.1 and Figure 3.10). General systems theory was used as the theoretical lens for this study (Figure 2.11). Based upon general system theory, the theoretical conceptual framework from the literature (Figure 2.12) has three distinct stages: User, IRS, and Evaluation. Using this framework, derived from key concepts and ideas from the literature, Section A therefore presents:

- i) the design of the hybrid index,
- ii) the results and analysis of the three pilot tests, from which the information was drawn (this information is discussed in Volume II Appendices A, B and C),
- iii) the final build of the hybrid index,
- iv) a summary of the hybrid index design findings, and
- v) the research question is answered.

Answers for the first research question are now explored (section 1.4 and section 3.1) as per the research design in Figure 3.8.

### **4.3 The design of the hybrid index**

The objective for this indexing and IRS design when searching for a document using a query was to ensure that word ordinality and word proximity were maintained. The design of IRS-H is presented using the hybrid indexing method consisting of three artefacts, the IRS itself, and a pair of hybrid indices: the hybrid token index and the hybrid query index (Figure 4.2). In this design, when matching a query to a document, IRS-H utilises a pair of indices where the hybrid query index interrogates the hybrid token index and returns a result (Appendix A, sections A.2 and A.3).

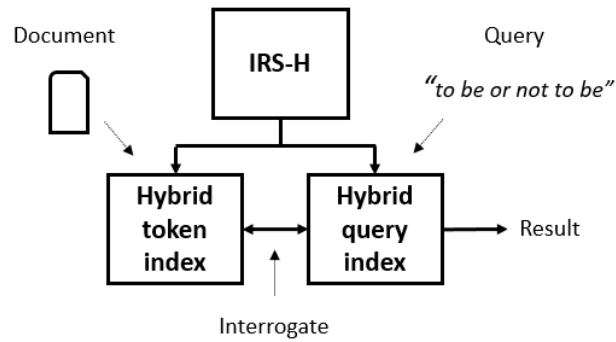


Figure 4.2: The three artefacts of IRS-H

To compare results from IRS-H, the IRS using the inverted index (IRS-I) was used as a control system (section 3.5.2). To put the design of IRS-I into perspective the system consists of two artefacts, the IRS itself, and an inverted index (Figure 4.3). When matching of a query to a document, IRS-I utilises the same singular inverted index and returns a result (Appendix A, section A.3).

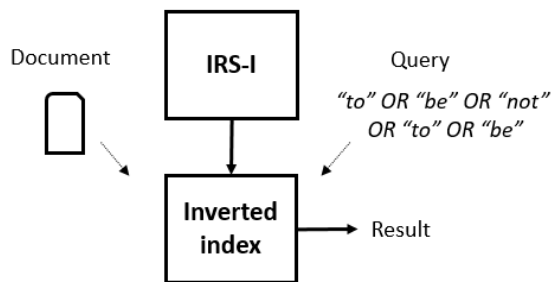


Figure 4.3: The two artefacts of IRS-I

Returning to the hybrid index, the design and functionality of the two indices (Figure 4.4) are now presented:

- i) During the IRSs information gathering phase, the hybrid token index is populated with each token acquired from the text of a document. Each token is sequentially allocated a unique Token ID that indicates the position of the token within the text of the document. The token, the document number, and the Token ID are stored in the hybrid token index.
- ii) In this example for the hybrid query index, the phrase-term '*to be or not to be*' is expressed within a query and is presented to the hybrid query index via the search engine. The number of words within the phrase-term is calculated and thereafter the hybrid token index is interrogated by the hybrid query index. The Token IDs for the first word and the last word are retrieved from the hybrid token index. If the numerical difference between the Token ID of the first word and the last word is equal, then the Boolean Match indicator is set to true. Match is set to true as the phrase-term '*to be or not to be*' is found to exist in the text of the document, the book Hamlet.

- iii) The hybrid query index is thus populated with each multi-word phrase-term expressed within a query, together with the document number, the begin Token ID representing the position of the first word in the phrase, the end Token ID representing the position of the last word in the phrase, and the Match indicator. When Match is true, the document is retrieved by the IRS for the user.

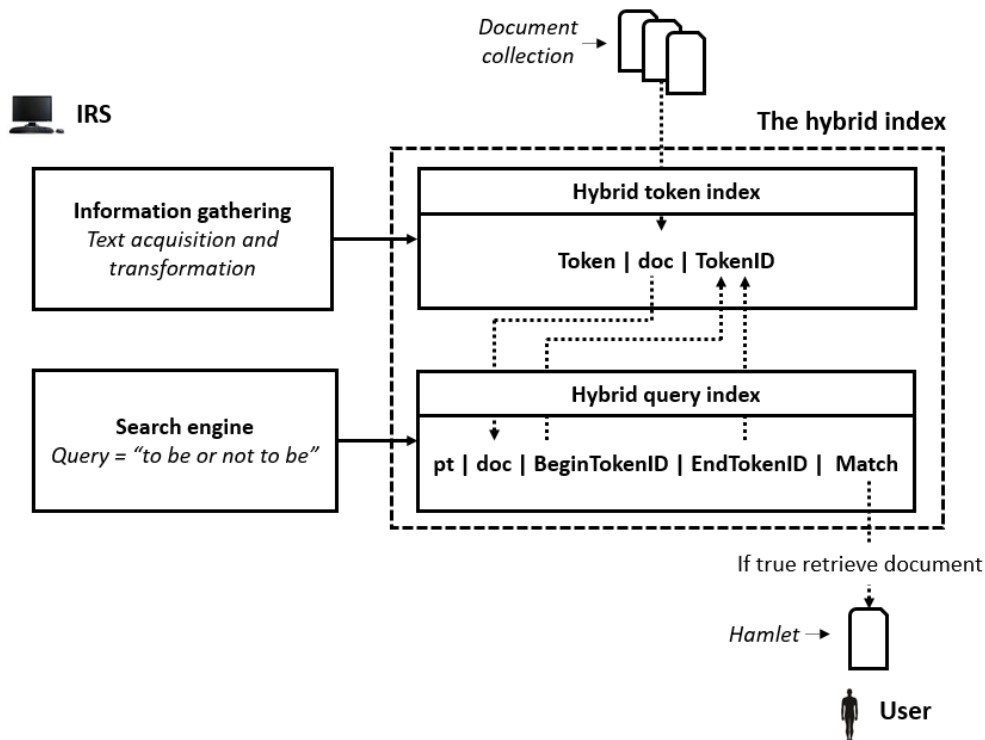


Figure 4.4: The design of the hybrid index

For further detail on the design of both indices, refer to Appendix A, sections A.2 and A.3.

#### 4.4 The pilot tests

The important design concepts, findings, and remedies from the three pilot tests (Appendices A, B and C) are now summarised:

##### 4.4.1 Pilot 1: Hamlet

Referring to the IRS in stage 2, extracted and replicated from the theoretical conceptual framework (Figure 2.12), the new design concepts (Figure 4.5) applied to Pilot 1 during the information gathering process are presented (Appendix A, section A.6):

- i) Pilot 1 was based on the book Hamlet Act 3 Scene 1 written by William Shakespeare circa 1599 (Shakespeare, 2018).

- ii) The inverted index was replaced by the pair of hybrid indices: the hybrid token index and the hybrid query index (Figure 4.2).
- iii) Content acquisition: the content from the single two-page document from Hamlet Act 3 Scene 1 was acquired from the pdf document and converted to text successfully.
- iv) Text transformation: text was case folded to lowercase, special characters were removed, and the tokens of text identified between delimiters were tokenised successfully.
- v) Hybrid token index: during pre-hybrid token index population, the document numbers and unique token IDs were allocated successfully. Thereafter the hybrid token index was populated with the tokens, document numbers, and unique token IDs.

The new design concepts applied to Pilot 1 during the search engine process were:

- i) Phrase-term: four phrase-terms provided by the user (the researcher in this pilot) were presented correctly, all in lowercase without special characters.
- ii) Phrase-term query: these phrase-terms were expressed as four queries, three singular and one expanded query.
- iii) Hybrid query index: the four queries were presented to the hybrid query index, which was thereafter populated with the phrase-terms, and unique begin and end token IDs.
- iv) The hybrid query index interrogated the hybrid token index successfully and where a match was found (a phrase-term existed in a document) the document number was returned and the hybrid query index updated accordingly.

The key findings from the design used in Pilot 1 were:

- i) Phrase-term frequency (*ptf*) needed to replace term frequency (*tf*) as by design, it was the number of phrase-terms that were required to be calculated rather than the single terms used in the inverted indexing method.
- ii) Converting *ptf* values to binary and the population of the phrase-term-by-document matrix (rather than the term-by-document matrix used in the inverted indexing method) with these values, was successful.
- iii) Stopping, the removal of stop words, the use of stemming, classifiers and suffix stripping were needless in this design. The tokens were not to be changed in any way thus preventing an in-exact match.
- iv) The IRS was able to match phrase-terms expressed in queries, held within the hybrid query index, to phrase-terms within the text of document held within the hybrid token index, exactly.

- v) Performance measurements were unusable as the judgment results from the user were unavailable and therefore not tested.
- vi) The sequentially generated number ranges made by the IRS were limiting. This applied to the document number and the index's unique token ID. To remedy this issue the number ranges were expanded accordingly.
- vii) The document length was a limiting factor that disallowed any benefit IRS-H may have had over IRS-I and vice versa. To remedy this issue the document length was increased.
- viii) At this stage, at the end of Pilot 1 there was no evidence to suggest that the functionality of IRS-H was more effective than IRS-I or vice versa.

In summary, the design changes to the key concepts of the theoretical conceptual framework for Pilot 1, highlighted in black, are presented in Figure 4.5.

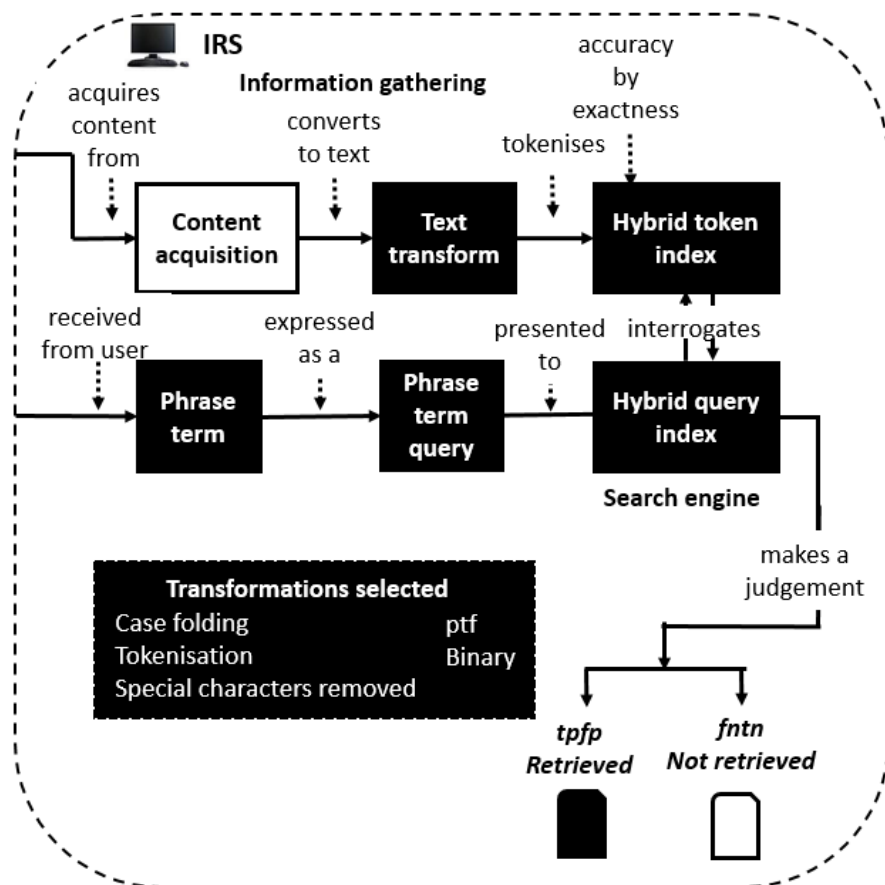


Figure 4.5: Pilot 1 design changes

#### 4.4.2 Pilot 2: Ulysses

The new design concepts (Figure 4.6) applied to Pilot 2 during the information gathering process were (Appendix B, section B.2):

- i) Pilot 2 was based on the book 'Ulysses', written by James Joyce (1932).

- ii) Content acquisition: document length was increased by altering the content to a single 666-page document, the book '*Ulysses*'. The content was acquired from the pdf document and converted to text successfully. However, on a few occasions the text was converted incorrectly by the OCR software.
- iii) Hybrid token index: the number ranges for the document number and the token ID were expanded as these had been limiting factors in Pilot 1. In addition, the token field in the index was expanded to accommodate larger sized tokens. For example, the token:  
*'handsomemarriedwomanrubbedagainstwidebehindinlonskeatram'*.

The new design concepts applied to Pilot 2 during the search engine process were:

- i) Phrase-term: 26 phrase-terms provided by the user (the researcher in this pilot) were presented correctly: all in lowercase without special characters.
- ii) Phrase-term query: these phrase-terms were expressed as 26 queries, six of which used single word phrase-terms.
- iii) Hybrid query index: the 26 queries were presented to the hybrid query index, which was thereafter populated with the phrase-terms and unique begin and end token IDs.
- iv) The hybrid query index interrogated the hybrid token index successfully and where a match was found (a phrase-term existed in a document) the document number was returned and the hybrid query index updated with the document number accordingly.

The key findings from the design used in Pilot 2 were:

- i) Phrase-term frequency (*ptf*) was maintained.
- ii) Converting *ptf* values to binary and the population of the phrase-term-by-document matrix with these values remained successful.
- iii) The IRS was able to match phrase-terms expressed in queries to those in documents exactly.
- iv) Performance measurements remained unusable, as the judgment results from the user were unavailable and therefore not tested.
- v) At this stage, at the end of Pilot 2 there was evidence to suggest that the functionality of IRS-H was more effective than IRS-I but this needed further investigation, testing, and evaluation with input from participating users.

In summary, the design changes to the key concepts for the theoretical conceptual framework for Pilot 2, highlighted in black, are presented in Figure 4.6.

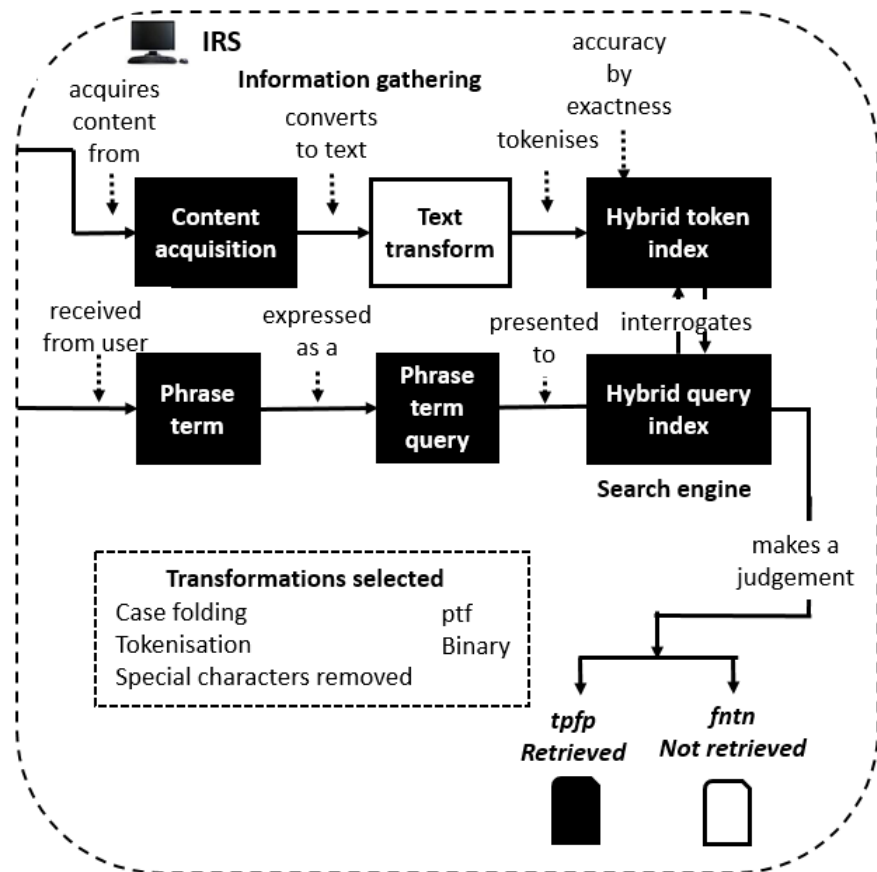


Figure 4.6: Pilot 2 design changes

#### 4.4.3 Pilot 3: Vocabulary mismatch

The new design concepts applied (Figure 4.7) to Pilot 3 during the information gathering process, are presented (Appendix C, section C.2):

- i) Pilot 3 was based upon a sample of 20 journal articles, conference papers and theses.
- ii) Content acquisition: the document collection was increased from a single document to 20 documents. The contents acquired from the pdf documents were converted to text successfully.

The new design concepts applied to Pilot 3 during the search engine process were:

- i) Phrase-term: 14 phrase-terms provided by the user (the researcher in this pilot) were presented correctly: all in lowercase without special characters.
- ii) Phrase-term query: these phrase-terms were expressed as 14 queries four of which were expanded queries.
- iii) Hybrid query index: the 14 queries were presented to the hybrid query index, which was thereafter populated with the phrase-terms and unique begin and end token IDs.

- iv) The hybrid query index interrogated the hybrid token index successfully and where a match was found (a phrase-term existed in a document) the document number was returned and the hybrid query index updated with the document number accordingly.

The key findings from the design used in Pilot 3 were:

- i) Phrase-term frequency (*ptf*) was maintained.
- ii) Converting *ptf* values to binary and the population of the phrase-term-by-document matrix with these values remained successful.
- iii) IRS-H was able to match phrase-terms expressed in queries to those in documents exactly.
- iv) IRS-H was able to maintain word ordinality and word proximity.
- v) At this stage, at the end of Pilot 3 there was evidence to suggest that the functionality of IRS-H was more effective than IRS-I but this needed further investigation, testing, and evaluation with input from participating users.

In summary, the design changes to the key concepts of the theoretical conceptual framework for Pilot 3, highlighted in black, are presented in Figure 4.7.

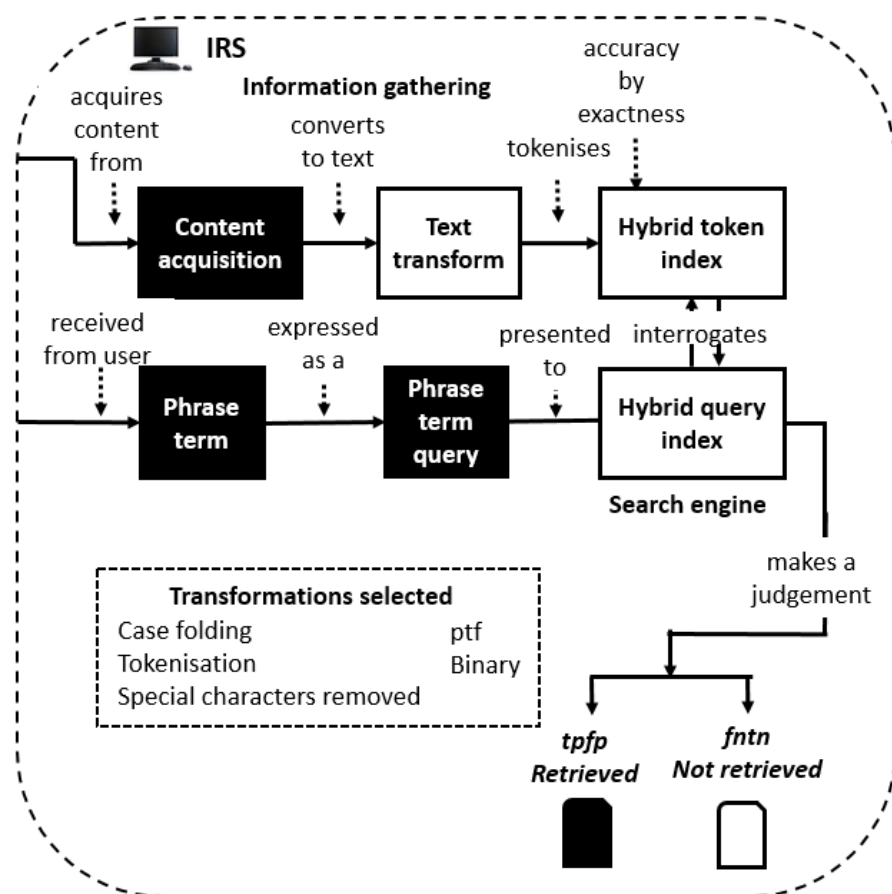


Figure 4.7: Pilot 3 design changes



## 4.5 Designing and building the hybrid index

The first research question asks how an IRS index can be designed that will maintain word ordinality and word proximity. To enable the research question to be answered, the build of the hybrid index is presented in two parts: i) to discuss the build of the hybrid token index as part of the IRSs information gathering process; and ii) to discuss the build of the hybrid query index as part of the IRSs search engine process.

Note that data used to explain concepts (Table 4.7) in this section 4.5 are derived from the design of the actual experiment (section 3.5.2) using 100 documents and the actual results of the experiment discussed in Chapter Four, Section B.

### 4.5.1 The information gathering process

Based on the theoretical conceptual framework (section 2.11, Figure 2.12), and now adapted by this design, the information gathering process illustrated in Figure 4.8 consists of three stages: i) text acquisition; ii) text transformation; and iii) the hybrid token index.

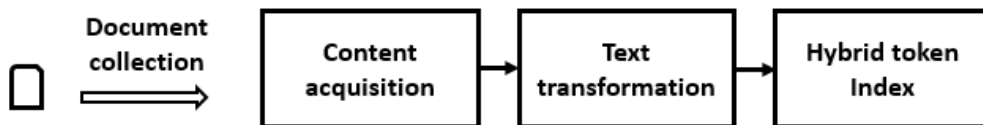


Figure 4.8: The information gathering process

#### 4.5.1.1 Content acquisition and text transformation

This section deals with both content acquisition and text transformation. Note that the document, ‘*A design science research methodology for information systems research*’ by Peffers et al. (2007) was drawn for the collection and used as an example.

There are three steps for content acquisition and text transformation illustrated in Figure 4.9: i) the original pdf document is added to the collection; ii) the PDF is converted to text format using OCR software known as content acquisition; and iii) IRS-H transforms the text file into a second text file known as text transformation. Note that these words (which become tokens) in the original text are case folded, special characters are removed and the tokens are pipe ‘|’ delimited. In addition, the tokens are sequentially ordered and this order is maintained when the hybrid token index is populated (for a full explanation of text transformation process, refer to Appendix A, section A.2.1.2).

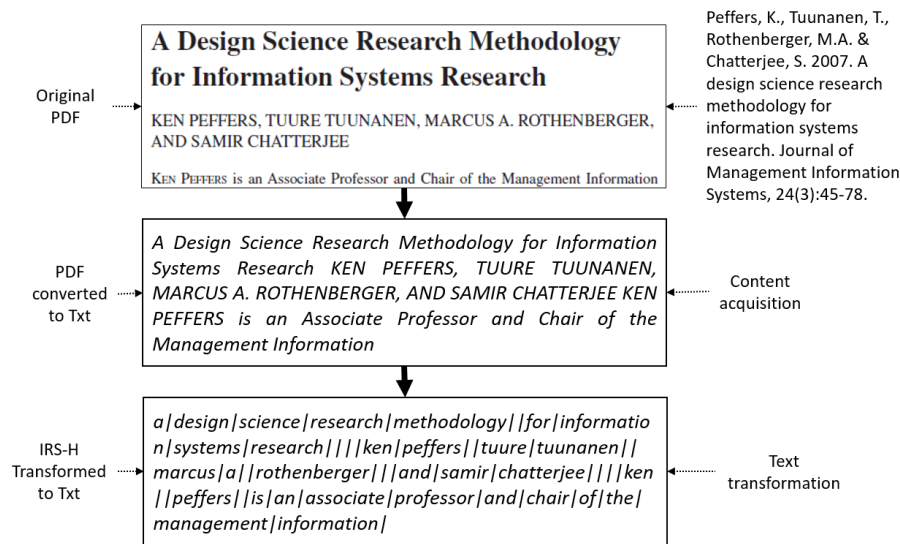


Figure 4.9: An example of content acquisition and text transformation

#### 4.5.1.2 The hybrid token index

The third stage of the information gathering process is the hybrid token index (Figure 4.2 and Figure 4.4). This index is represented by a data base table designed with three columns. The first column stores a non-distinct list of tokens acquired from the transformed text during the information gathering process (embedded in the theoretical conceptual framework section 2.11, Figure 2.12 and updated in Figure 4.7).

The second column stores a single document number indicating where the token exists within the text of that specific document. The third column stores the unique Token ID, a sequential system-generated number unique to that token within that document within the collection. As the fully populated index is far too voluminous to present in this thesis, it stores 983,081 tokens, an example using the first 30 tokens acquired from document  $d_{0002}$  (based on the work of Peffers et al., 2007) is presented for the hybrid token index in Table 4.2 (Figure 4.9 for the content).

Table 4.2: An example of the hybrid token index

Token	doc	TokenID	Token	doc	TokenID	Token	doc	TokenID
a	d0002	10008407	peffers	d0002	10008417	peffers	d0002	10008427
design	d0002	10008408	tuure	d0002	10008418	is	d0002	10008428
science	d0002	10008409	tuunanen	d0002	10008419	an	d0002	10008429
research	d0002	10008410	marcus	d0002	10008420	associate	d0002	10008430
methodology	d0002	10008411	a	d0002	10008421	professor	d0002	10008431
for	d0002	10008412	rothenberger	d0002	10008422	and	d0002	10008432
information	d0002	10008413	and	d0002	10008423	chair	d0002	10008433
systems	d0002	10008414	samir	d0002	10008424	of	d0002	10008434
research	d0002	10008415	chatterjee	d0002	10008425	the	d0002	10008435
ken	d0002	10008416	ken	d0002	10008426	management	d0002	10008436

#### 4.5.1.3 Token index entity relationship diagram

To facilitate the design of the hybrid token index, two data base tables were used. The first table is the document table that stores the data pertaining to each of the 100 documents in the collection. The second table is the hybrid token index table that stores the token indexing data discussed earlier (Table 4.2). There is a one-to-many (1 -  $\infty$ ) relationship between the document table and the hybrid token index table as illustrated in the entity relationship diagram (ERD) (Figure 4.10).

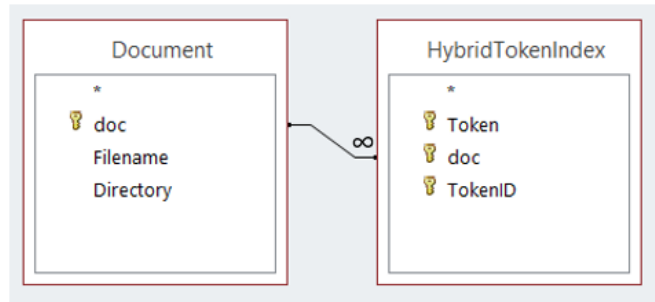


Figure 4.10: Entity relationship diagram for hybrid token indexing

Referring to Figure 4.10, the document table on the left contains three fields: the document number (doc), the physical file name (Filename) of the document, and the directory (Directory) or file path indicative of where the document resides on the computer. The full document table containing the 100 documents is presented in Appendix D, Table D.1. The hybrid token index on the right contains three fields: the token (Token), the document number (doc) and the unique token ID (TokenID) of the token.

This ends the design of the information gathering process for IRS-H. The next section explains the design of the search engine process using the hybrid query index.

#### 4.5.2 The search engine process

Based on the theoretical conceptual framework (section 2.11, Figure 2.12), and now adapted by this design, the search engine process, illustrated in Figure 4.11, consists of three stages: i) the phrase-term; ii) the phrase-term query; and iii) the hybrid query index.

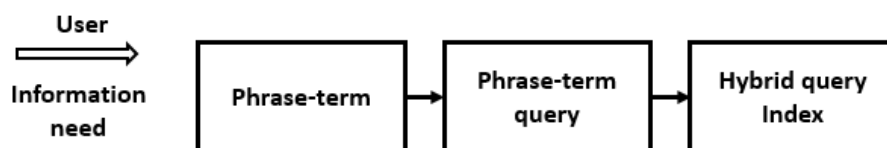


Figure 4.11: The search engine process

#### 4.5.2.1 The phrase-term

To enable a search to return documents to satisfy a user's information need, the user specifies specific phrase-terms. Once specified these multi-word phrase-terms, created by the user, are then gathered and stored with the phrase-term table (refer to the query index ERD later in this section). The full list of 75 information needs is presented in Appendix D, Table D.2, and the full list of 65 phrase-terms used in this research is presented in Appendix D, Table D.4.

#### 4.5.2.2 The phrase-term query

These phrase-terms are then expressed as queries. The goal of the search engine is to try to match the query to the document. Once specified these phrase-term queries, which have been created by the user, are then gathered and stored with the phrase-term table (refer to the query index ERD later in this section). The full list of 75 phrase-term queries used in this research is presented in Appendix D, Table D.3.

#### 4.5.2.3 The hybrid query index

A second index, referred to as the hybrid query index (Figure 4.2 and Figure 4.3), is designed to hold the structure of the phrase-terms expressed within the queries. It is a five-column data base table positioned within the IRS data store (in Table 4.3). The first column stores the phrase-term number, the second column stores the document number, the third and fourth columns (begin Token ID and end Token ID respectively) store the position of the words within the phrase-term to maintain word ordinality and proximity, and the fifth column is a Boolean flag to indicate whether a match has been found (between the phrase-term expressed within the queries and the text of the document) or not. When a match occurs, the index flag (Match) is set to true and the document numbers are returned by the IRS as *tpfp*, defined as the number of documents retrieved by the IRS. As the fully populated index is far too voluminous to present in this thesis – the index stores 1,274,286 possible outcomes – an example of the hybrid query index presenting the results of the first five of the 28 terms is illustrated in Table 4.3.

Table 4.3: An example of the hybrid query index

pt	doc	Begin Token ID	End Token ID	Match
pt01	d0002	10008408	10008409	True
pt01	d0002	10008642	10008643	True
pt01	d0002	10008725	10008726	True
pt01	d0002	10009032	10009033	True
pt01	d0002	10009073	10009074	True

Referring to the process flow (embedded in the theoretical conceptual framework section 2.11, Figure 2.12 and updated in Figure 5.4) of IRS-H, the hybrid indexing method is as follows: the information gathering process is followed and the hybrid token index is populated with the tokens, related document numbers, and the unique Token ID. Words are then expressed as single or multi-word phrase-terms in one or more queries in an attempt to represent a user's information needs. The search engine then stores the structural information of each phrase-term (each word, its position, and its sequence) in the hybrid query index. The search engine then presents these queries from the query index to the hybrid token index (Figure 4.2) and attempts to match, through interrogation, the words within the phrase-terms expressed within the queries to the tokens within the hybrid token index, simultaneously maintaining word ordinality and proximity. When a match is found IRS-H returns the document numbers for each of the phrase-terms and these numbers are then stored in a phrase-term-by-document matrix within IRS-H.

#### **4.5.2.4 Query index entity relationship diagram**

To facilitate the design of the hybrid query index, seven data base tables were used:

- i) the information need table stores the information need number and description of each information need,
- ii) the query table stores the query number and description of each query,
- iii) the phrase-term table stores the phrase-term number and the words used in each phrase-term,
- iv) the information need query linking table stores the one-to-one relationships of information need to query, using the information need number and the query number respectively,
- v) the query phrase-term linking table stores the one-to-many (1 -  $\infty$ ) relationships of query to phrase-term, using the query number and the phrase-term number respectively,
- vi) the hybrid query index table stores the query indexing data based on the query's phrase-terms discussed earlier, and
- vii) the document table used during the hybrid token indexing process.

The ERD illustrating these seven tables and their relationships is presented in Figure 4.12.

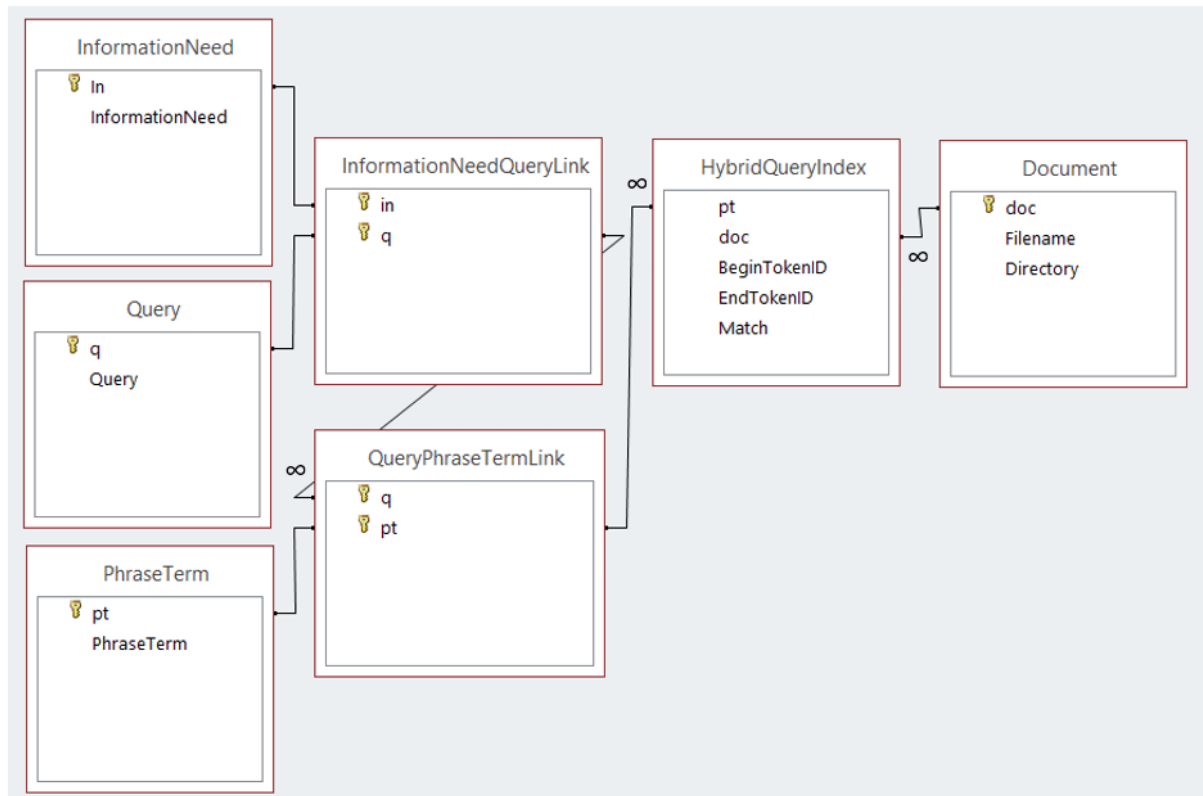


Figure 4.12: Entity relationship diagram for hybrid query indexing

The design of the fields and their relationships within the tables of the ERD are further discussed:

- i) InformationNeed table – the information need table at the top left contains two fields: the information need number (*In*) and the full description of the information need (InformationNeed).
- ii) Query table – below the information need table is the query table that contains two fields: the query number (*q*) and the full structure of the query (Query).
- iii) PhraseTerm table – below the query table is the phrase-term table that contains two fields: the phrase-term number (*pt*) and the full multi-word phrase term (PhraseTerm).
- iv) InformationNeedQueryLink table – at the centre left are two tables. The first is the Information need query link table (that links an information need to a query) and contains two fields: the information need number (*In*) and the query number (*q*).
- v) QueryPhraseTermLink table – the second is the Query phrase-term link table (that links a query to a phrase-term) and contains two fields: the query number (*q*) and the phrase-term number (*pt*).
- vi) HybridQueryIndex table – at the centre right is the hybrid query index table that contains five fields: the phrase-term number (*pt*), the document number

(doc), the 'begin token ID' number (BeginTokenID), the 'end token ID' number (EndTokenID), and the Boolean match indicator (Match).

- vii) Document table – the final table on the right is the document table (Figure 4.12) that contains three fields: the document number (doc), the physical file name (Filename) of the document and the directory (Directory) or file path of where the document resides on the computer.

#### 4.6 Summary of hybrid indexing design findings

The summary of the findings for the design of the hybrid token index, the hybrid query index, and the results from the three pilot tests are presented in Table 4.4.

Table 4.4: Summary of hybrid indexing design findings

No	Finding
1	The inverted index was replaced by the pair of hybrid indices, the hybrid token index and the hybrid query index, and functioned successfully.
2	For content acquisition, document length was initially a limiting factor but this was remedied with the IRS accommodating at least 20 documents. The conversion from pdf to text was successfully except for a few incorrect tokens made by the OCR software in Pilot 2.
3	For text transformation, text was case folded to lowercase, special characters were removed, and the tokens of text identified between delimiters were tokenised successfully. Stopping, the use of stemming, classifiers, and suffix stripping were needless in this design.
4	The hybrid token index was able to handle the tokens acquired from the text and store them in a way that maintained word ordinality and word proximity. Number ranges for document numbers and Token IDs were increased and the token field size was expanded to cater for large sized tokens.
5	Phrase-terms presented in lowercase without special characters were handled successfully by the IRS.
6	Phrase-term queries containing single word or multi-word phrase-terms together with expanded queries were handled successfully by the IRS.
7	The hybrid query index was able to handle the phrase-terms expressed within the queries and store them in a way that maintained their 'begin' and 'end' positions. Number ranges for document numbers and Token IDs were initially limited and later increased.
8	The hybrid query index interrogated the hybrid token index successfully and where a match was found (a phrase-term existed in a document) the document number was returned and the hybrid query index was updated accordingly.
9	The IRS was able to match phrase-terms expressed in queries, held within the hybrid query index, to phrase-terms within the text of document held within the hybrid token index, exactly.
10	Phrase-term frequency ( <i>ptf</i> ) replaced term frequency ( <i>tf</i> ) as by design, it was the number of phrase-terms that were required to be calculated rather than single terms used in the inverted indexing method. As <i>df</i> and <i>idf</i> can be derived from <i>tf</i> , and <i>tf_idf</i> from <i>tf</i> and <i>idf</i> , and because these values were not required because of exact matching, these values were not catered for within the IRS.
11	Converting <i>ptf</i> values to binary and the population of the phrase-term-by-document matrix with these values was successful.
12	During these pilot tests, performance measurements were unusable as the judgment results from the user were unavailable and therefore not tested.
13	At this stage of the research there was no evidence to suggest that the functionality of IRS-H was more effective than IRS-I or vice versa.

#### **4.7 The first research question**

To complete Section A, the first research question is now addressed. The research question is:

##### **RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

In summary, the hybrid indexing design utilises a pair of hybrid indices: the hybrid token index and the hybrid query index, using the concept of the unique Token ID. The objective in the design was to find an indexing method that maintained word ordinality and word proximity. The design of the hybrid indexing method is one way an IRS can be designed that will maintain word ordinality and word proximity. During the information gathering process, the hybrid token index of IRS-H is populated with tokens acquired from the text and each is allocated a unique Token ID. This Token ID is the key concept in this design and maintains the ordinality of the words and the proximity of words. When the search engine process is activated, the words within the query's multi-word phrase-terms are matched exactly to the tokens in the hybrid token index with correct word ordinality and proximity.

However, at this stage it cannot be established which one of the two methods (IRS-H or IRS-I) is more effective: the traditional inverted indexing method or the hybrid indexing method, thus the reasoning behind the explanatory study using experimentation and evaluation, of which the results are presented in Section B of this chapter.

### **SECTION B – THE EXPLANATORY STUDY**

#### **4.8 The experiment – an introduction**

Section A was the exploratory study using DSR to design and build IRS-H and its pair of hybrid indices. Section B is the explanatory study based on the experiment described in the flow chart in section 3.3.3, Figure 3.4. General systems theory was used as the theoretical lens for this study (Figure 2.11). Based upon general system theory, the theoretical conceptual framework from the literature (Figure 2.12) with its three distinct stages of User, IRS, and Evaluation, was used as a framework to present Section B. Section B is the analytical quantitative section of this research that tests the five hypotheses and presents:

- i) the experiment (both user and systems),
- ii) the results of analysing effectiveness in retrieving relevant documents,
- iii) the results of analysing incorrect identification of relevant documents,
- iv) the results of analysing rejection quality of non-relevant documents,



- v) the results in solving the vocabulary mismatch problem,
- vi) the results in satisfying the user's information need, and
- vii) a summary of the experimental findings.

## 4.9 The experiment

The experiment was based on the experimental framework described in Chapter Three, section 3.6.3 and Figure 3.16. This framework is re-drawn for the benefit of the reader in Figure 4.13 and is discussed.

### 4.9.1 The User

Using a collection of 100 documents (Appendix D, Table D.1), a set of 75 information needs (Appendix D, Table D.2), and 75 queries (Appendix D, Table D.3) using 65 phrase-terms (Appendix D Table D.4) were presented to the five participants during the user experiment, in the form of a questionnaire (Appendix E). Thereafter the data from the questionnaire were collected and arranged in: i) an information-need-by-document matrix (Appendix F, Table F.1); ii) a phrase-term-by-document matrix (Appendix D, Table F.2); and iii) a query-by-document matrix that contained the binary values derived from ii).

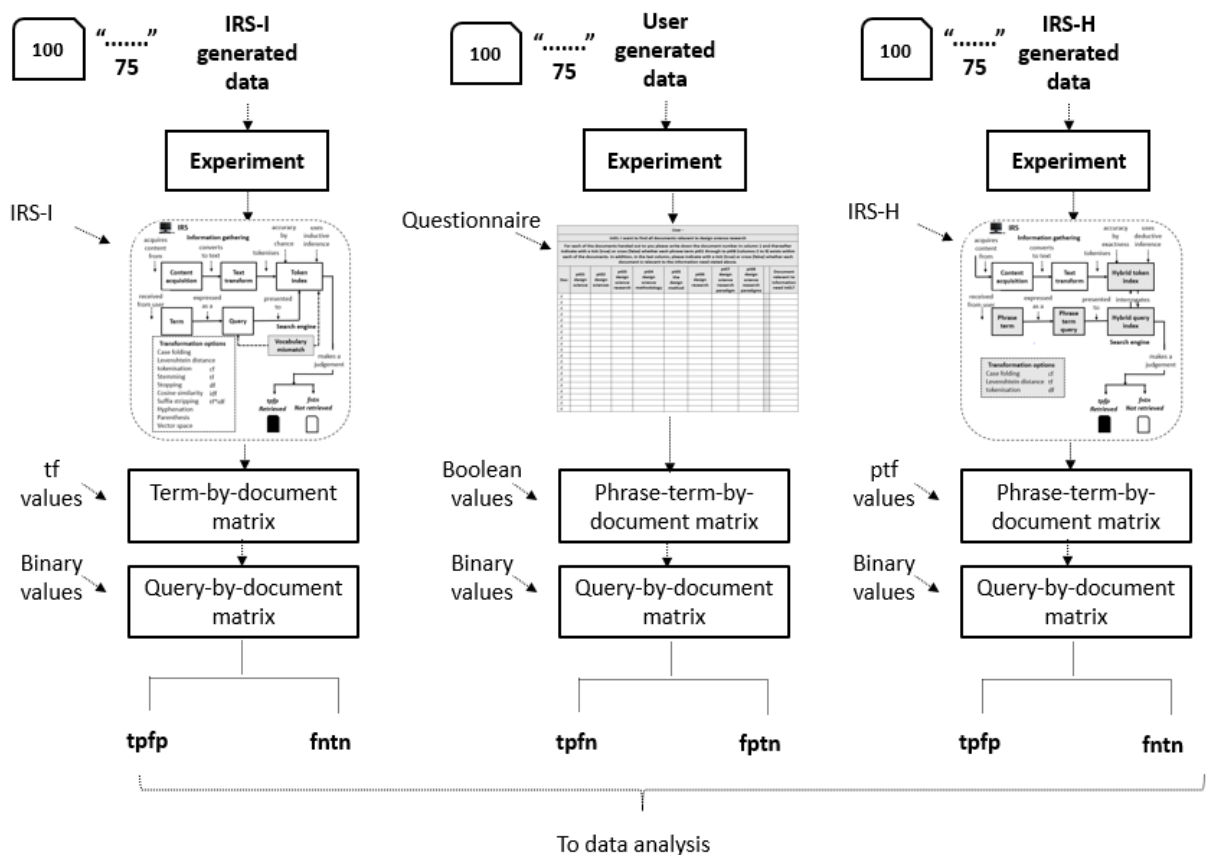


Figure 4.13: The experimental framework

#### 4.9.2 IRS-H

The same sets of documents, information needs, phrase-terms, and phrase-term queries were presented to IRS-H, and after search engine activation, these generated system data based on the hybrid indexing method. The data generated were arranged in: i) an information-need-by-document matrix (Appendix G, Table G.1); ii) a phrase-term-by-document matrix (Appendix G, Table G.2); and iii) a query-by-document matrix that contained the binary values derived from ii).

#### 4.9.3 IRS-I

The process for IRS-H was repeated for IRS-I. However, the set of queries were adapted to suit the inverted index method format. The same sets of documents and information needs were used, however the 65 phrase-terms were replaced with 49 single-word terms, and the 75 queries were rearranged to accommodate these terms. The terms and term-queries were presented to IRS-I and after search engine activation, these generated system data based on the inverted indexing method. The data generated was arranged in: i) an information-need-by-document matrix (Appendix H, Table H.1), ii) a term-by-document matrix (Appendix H, Table H.2); and iii) a query-by-document matrix that contained the binary values derived from ii).

During the experiment, data were generated by running IRS-H together with the user data to produce specific key values. This generation of data was repeated for IRS-I, thus producing two sets of data allowing IRS-H to be compared with IRS-I. The key values were:  $tp$ ,  $tn$ ,  $fp$ ,  $fn$  and  $tf$  which could now be applied to the various formulae that produce the performance measurements of Precision, Recall, F-measure, Specificity, the 2x2 contingency table, the term-by-document matrix and  $tp$ ,  $tn$ ,  $fp$ ,  $fn$  and  $tf$  (These performance measurements were then used to test the five hypotheses of this study (section 3.5.2.3 and Figure 3.13).

At this stage the research hypotheses were tested (section 1.4 and section 3.1) as per the research design in Figure 3.8. Testing hypotheses **H1**, **H2** and **H3** follow the IRS-H versus IRS-I system-generated flow chart described in section 3.5.2.3, Figure 3.13.

#### 4.10 Hypothesis 1: Analysing effectiveness in retrieving relevant documents

For the first hypothesis **H1**, the null and alternative hypotheses are stated as:

**H1<sub>0</sub>**: Hybridised indexing does not increase the effectiveness of retrieving relevant documents

**H1<sub>1</sub>:** Hybridised indexing increases the effectiveness of retrieving relevant documents

To test the null hypothesis for the first hypothesis, the calculations for the mean average precision (MAP) for both indexing methods were required to determine whether or not IRS-H increases the effectiveness of the retrieval of user relevant documents compared to IRS-I. The experimental results are now presented as follows: i) the precision formula; ii) the precision contingency table logic; iii) examples of ranking to calculate average precision per query per indexing method; iv) the results for all ranked average precision measurements; v) the MAP formula with examples; and vi) finally the student's t-test and t-distribution results.

#### 4.10.1 Precision measurements

Precision can be expressed as in equation 4.1 (Manning et al., 2008:143; Narayan et al., 2017:2).

$$P = \frac{tp}{tp + fp}$$

Equation 4.1: P (Manning et al., 2008:143; Narayan et al., 2017:2)

$tp$  and  $fp$  represent IRS retrieved documents (user relevant and user non-relevant respectively), while  $tp$  represents user relevant IRS retrieved documents. The result for precision (refer to the performance measurements in Appendix I) for IRS-I query  $q_{01}$  was calculated as follows:

$$P_{\text{IRS-I}q_{01}} = \frac{23}{23 + 49} = 0.32$$

Equation 4.2: PIRS-Iq01

Similarly, for IRS-H, the first row represents query  $q_{01}$  and the result for precision was calculated as follows:

$$P_{\text{IRS-H}q_{01}} = \frac{13}{13 + 9} = 0.59$$

Equation 4.3: PIRS-Hq01

#### 4.10.2 Ranking

One table was created for each indexing method to store the average precision measurements per query. This resulted in 7,500 records, as there were 75 queries multiplied by 100 documents. As these tables were too large to present in this thesis,

two examples are now provided of the method used to determine the average precision measurements. The calculation of average precision (AP) was based on the precision values of ranked positions from which relevant and non-relevant documents were retrieved for each query. If no documents were retrieved by the IRS for a query, the query's AP was set to zero.

#### 4.10.2.1 IRS-I Ranking

In the first example, the precision ranking method for IRS-I query  $q_{01}$  is explained. Out of a collection of 100 documents, query  $q_{01}$  retrieved 72 documents of which 23 were judged relevant by the user. These documents are presented in Figure 4.14.

At rank position 1, the precision value is the number of relevant documents divided by the number of user relevant retrieved documents. In this case, precision at rank position 1 is  $0 / 1 = 0$ , at rank position 2 is  $(0 + 1) / (1 + 1) = 0.50$ , at rank position 3 is  $(0 + 1 + 0) / (1 + 1 + 1) = 0.33$ , etc. To calculate the average precision for the query, only the rankings for the relevant documents are summed and then divided by the number of user relevant documents, for example, Average precision for IRS-I $_{q_{01}}$  =  $((0.5 + 0.33 + 0.16 + 0.20 + 0.23 + 0.24 + 0.23 + 0.26 + 0.26 + 0.27 + 0.29 + 0.31 + 0.30 + 0.32 + 0.33 + 0.35 + 0.33 + 0.34 + 0.35 + 0.36 + 0.38 + 0.39 + 0.32) / 23) = 0.31$ . The results for all ranked average precision measurement are presented in Table 4.5 for both indexing methods.

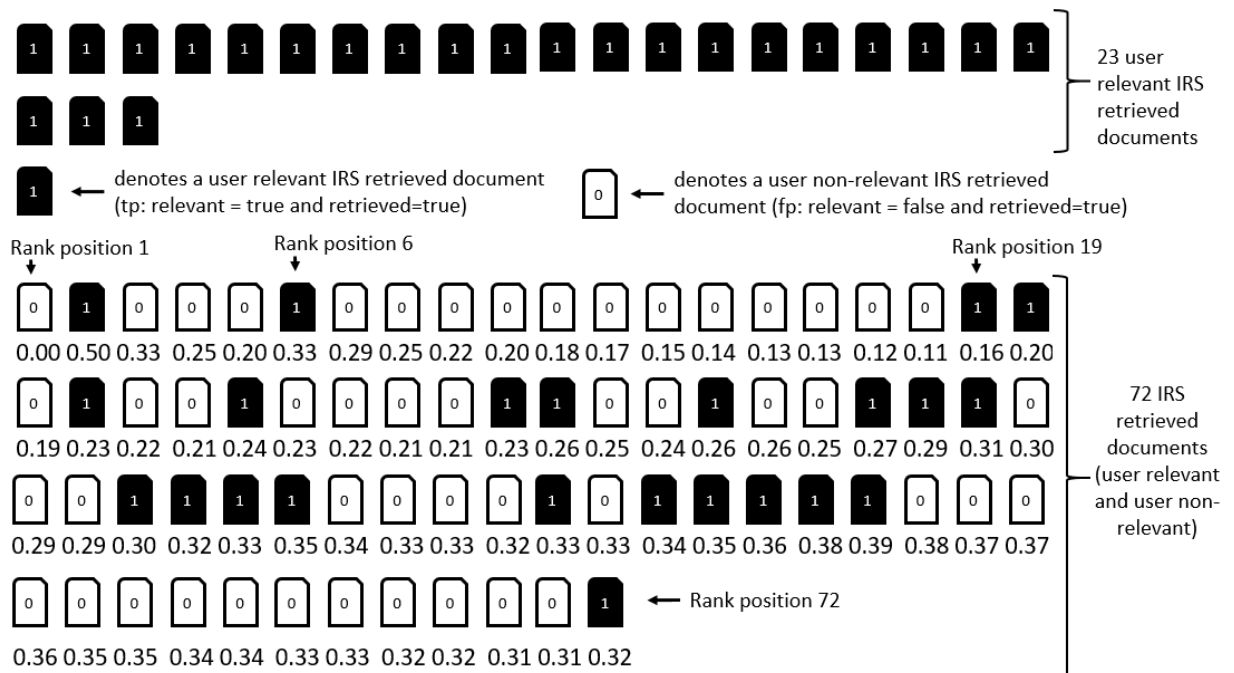


Figure 4.14: IRS-I query 1 – Precision values for one ranking of 23 relevant documents

#### 4.10.2.2 IRS-H Ranking

In the second example, the precision ranking method for IRS-H query  $q_{01}$  is explained. Out of a collection of 100 documents, query  $q_{01}$  retrieved 22 documents of which 13 were judged relevant by the user. These documents are presented in Figure 4.15.

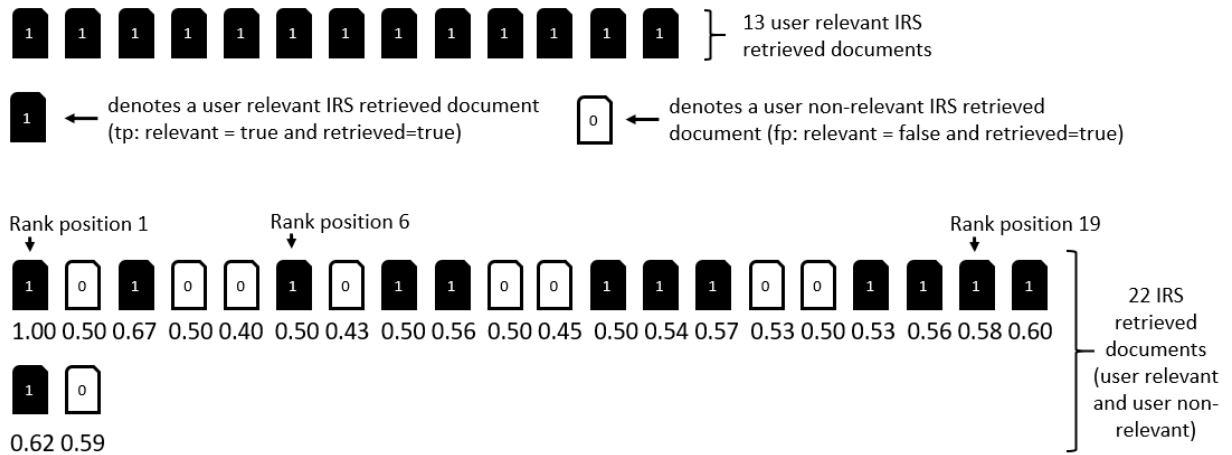


Figure 4.15: IRS-H query 1 – Precision values for one ranking of 13 relevant documents

At rank position 1, the precision value is the number of relevant documents divided by the number of user relevant retrieved documents. In this case, precision at rank position 1 is  $1 / 1 = 1$ , at rank position 2 is  $(1 + 0) / (1 + 1) = 0.50$ , at rank position 3 is  $(1 + 0 + 1) / (1 + 1 + 1) = 0.67$ , etc.

To calculate the average precision for the query, only the rankings for the relevant documents are summed and then divided by the number of user relevant documents, for example, Average precision for IRS-H $_{q_{01}}$  =  $((1.00 + 0.67 + 0.50 + 0.50 + 0.56 + 0.50 + 0.54 + 0.57 + 0.53 + 0.56 + 0.58 + 0.60 + 0.62) / 13) = 0.59$ .

#### 4.10.3 Average precision measurements

The results for all ranked average precision measurements are presented in Table 4.5 for both indexing methods.

Table 4.5: Results of average precision measurements per query per indexing method

Query	IRS-I Average Precision	IRS-H Average Precision	Query	IRS-I Average Precision	IRS-H Average Precision	Query	IRS-I Average Precision	IRS-H Average Precision
q01	0.31	0.59	q26	0.24	0	q51	0.23	0
q02	0.29	0.57	q27	0.24	0	q52	0.23	0.52
q03	0.27	0.62	q28	0	0	q53	0.23	0.52
q04	0.27	0	q29	0	0	q54	0.23	0.52
q05	0.27	0.68	q30	0	0	q55	0.23	0
q06	0.27	0.71	q31	0	0	q56	0.23	0
q07	0.27	0	q32	0	0	q57	0.23	0
q08	0.26	0	q33	0.23	0.55	q58	0.23	0.52

Query	IRS-I Average Precision	IRS-H Average Precision
q09	0.26	0
q10	0.26	0.7
q11	0.26	0.64
q12	0.25	0
q13	0.25	0
q14	0.24	0
q15	0.24	0
q16	0.23	0
q17	0.24	0.58
q18	0.23	0.58
q19	0.23	0
q20	0.23	0
q21	0.23	0
q22	0.23	0
q23	0.23	0.58
q24	0.23	0.58
q25	0.23	0

Query	IRS-I Average Precision	IRS-H Average Precision
q34	0.23	0.53
q35	0.23	0
q36	0.23	0
q37	0.23	0.52
q38	0.23	0.52
q39	0.23	0.52
q40	0.23	0
q41	0.23	0
q42	0.23	0.53
q43	0.23	0.53
q44	0.23	0.51
q45	0.23	0
q46	0.23	0
q47	0.23	0
q48	0	0
q49	0.23	0.51
q50	0.23	0.52

Query	IRS-I Average Precision	IRS-H Average Precision
q59	0.23	0
q60	0.23	0.53
q61	0.23	0.53
q62	0.23	0
q63	0.23	0
q64	0.23	0.54
q65	0.23	0.54
q66	0.23	0.54
q67	0.23	0.53
q68	0.23	0.52
q69	0.22	0.51
q70	0.22	0
q71	0.22	0.49
q72	0.22	0.49
q73	0.21	0.49
q74	0.22	0.5
q75	0.22	0.49

#### 4.10.4 Mean average precision

Mean average precision (MAP) can be expressed as the sum of the average precision values for all queries divided by the number of queries as presented in equation 4.4 (Zhao & Huang, 2016:3).

$$MAP = \frac{\sum_{q=1}^Q AveP(q)}{Q}$$

Equation 4.4: MAP (Zhao & Huang, 2016:3)

The calculated MAP results for all queries for IRS-I ( $MAP_{IRS-I}$ ) and IRS-H ( $MAP_{IRS-H}$ ) are:

$$MAP_{IRS-I} = \frac{16.34}{75} = 0.2179$$

Equation 4.5: MAP<sub>IRS-I</sub>

$$MAP_{IRS-H} = \frac{20.85}{75} = 0.2780$$

Equation 4.6: MAP<sub>IRS-H</sub>

As  $MAP_{IRS-I} = 0.2179$  and  $MAP_{IRS-H} = 0.2780$  then  $MAP_{IRS-H} > MAP_{IRS-I}$ . This suggests the mean average precision of IRS-H is greater than the mean average precision of IRS-I. To prove statistically that these results did not occur by chance, a student's t-test was performed.

#### 4.10.5 Student's t-test and t-distribution

A two-tailed student's t-test is traditionally used to test the difference between the means of two systems and to determine whether this difference is statistically significant (Smucker et al., 2007). For the first hypothesis, a one-tailed student's t-test was used to test whether the mean average precision of IRS-H ( $MAP_{IRS-H}$ ) was greater than the mean average precision of IRS-I ( $MAP_{IRS-I}$ ) and whether the result was statistically significant.

To perform the t-test, a 95% confidence level was used resulting in a significance level ( $\alpha$ ) of 5%. As each system had 75 values, one per query, the sample size ( $N$ ) was 150 and as there were two systems, the degrees of freedom ( $df$ ) equalled the sample size minus the number of systems (on the level of average) in the test. As these systems (averages) were  $MAP_{IRS-I}$  and  $MAP_{IRS-H}$ ,  $df = 150 - 2 = 148$ . The critical value ( $t_{cv}$ ) for the one-tailed t-test using a significance level of 0.05 and a  $df$  of 148 was 1.66. The MAP t-distribution and results are presented in Figure 4.16.

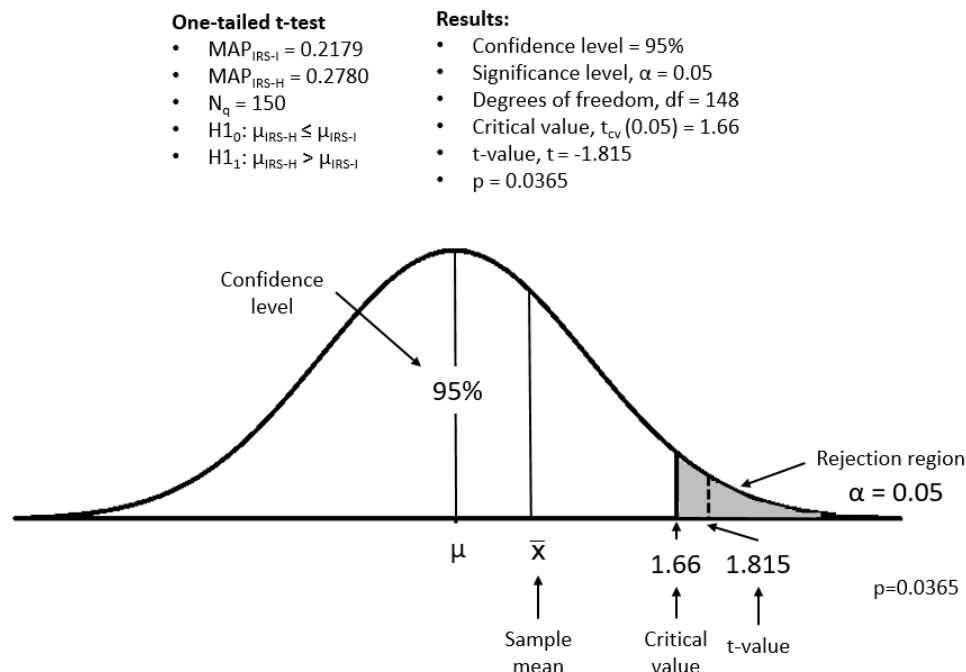


Figure 4.16: Mean average precision t-distribution and results

After performing the t-test, the t-value ( $t$ ) result was 1.815. To accept the null hypothesis  $H_{10}$  the t-value must be less than or equal to  $t_{cv} = 1.66$ . As  $t$  is not less

than or equal to  $t_{cv}$  (as  $t > t_{cv}$ ) the null hypothesis is rejected and therefore a Type I error was made as it is a rejection of the null hypothesis. The results are statistically significant as  $p < \alpha$  where  $p$  (Type I error) = 0.0365 and  $\alpha = 0.05$ .

As  $t > t_{cv}$ , where  $t = 1.815$  and  $t_{cv} = 1.6$  the alternative hypothesis **H1**<sub>1</sub> is accepted and therefore:

*Hybridised indexing increases the effectiveness of retrieving relevant documents.*

#### 4.11 Hypothesis 2: Analysing incorrect identification of relevant documents

For hypothesis **H2**, the null and alternative hypotheses are stated as:

**H2**<sub>0</sub>: Hybridised indexing does not reduce the incorrect identification of relevant documents

**H2**<sub>1</sub>: Hybridised indexing reduces the incorrect identification of relevant documents

To test the null hypothesis for the second hypothesis, the calculations for the mean average recall (MAR) for both indexing methods was required to determine whether or not IRS-H reduces the incorrect identification of user relevant documents, compared to IRS-I. Thereafter, to confirm the results were not by chance, a one-tailed student's t-test was performed to verify the MAR results. The experimental results for the second hypothesis are now presented as follows: i) the recall formula; ii) the recall contingency table logic; iii) examples of ranking to calculate average recall per query per indexing method; iv) the presentation of the results for all ranked average recall measurements; v) the MAR formula with examples; and vi) the student's t-test and t-distribution results, to determine whether the results are statistically significant or not.

##### 4.11.1 Recall measurements

Recall can be expressed as indicated in equation 4.7 (Manning et al., 2008:143; Narayan et al., 2017:2):

$$R = \frac{tp}{tp + fn}$$

**Equation 4.7: R (Manning et al., 2008:143; Narayan et al., 2017:2)**

$tp$  and  $fn$  represent user relevant documents (IRS retrieved and IRS not-retrieved respectively), while  $tp$  represents user relevant IRS retrieved documents. The result for Recall (refer to the performance measurements in Appendix I) for IRS-I query  $q_{01}$  was calculated as follows:



$$R_{\text{IRS-I}_{q01}} = \frac{23}{23 + 2} = 0.92$$

Equation 4.8: RIRS-Iq01

Similarly, for IRS-H, the first row represents query  $q_{01}$  and this result for Recall was calculated as follows:

$$R_{\text{IRS-H}_{q01}} = \frac{13}{13 + 12} = 0.52$$

Equation 4.9: RIRS-Hq01

#### 4.11.2 Ranking

One table was created for each indexing method to store the average recall measurements per query. This resulted in 7,500 records, as there were 75 queries multiplied by 100 documents. As these tables were too large to present in this thesis, two examples are now provided of the method used to determine the average recall measurements. The calculation of average recall (AR) was based on the recall values of the ranked positions from which relevant and non-relevant documents were retrieved for each query. If no documents were retrieved by the IRS for a query, the query's AR was set to zero.

##### 4.11.2.1 IRS-I ranking

In the first example, the recall ranking method for IRS-I query  $q_{01}$  is explained. Out of a collection of 100 documents, query  $q_{01}$  retrieved 72 documents of which 23 were judged relevant by the user. These documents are presented in Figure 4.17.

At rank position 1, the recall value is the number of relevant documents divided by the number of user relevant retrieved documents. In this case, precision at rank position 1 is  $0 / 23 = 0$ , at rank position 2 is  $(0 + 1) / 23 = 0.04$ , at rank position 3 is  $(0 + 1 + 0) / 23 = 0.04$ , etc. To calculate the average recall for the query, only the rankings for the relevant documents are summed and then divided by the number of user relevant documents, for example, Average recall for IRS-I $_{q01}$  =  $((0.04 + 0.09 + 0.13 + 0.17 + 0.22 + 0.26 + 0.30 + 0.35 + 0.39 + 0.43 + 0.48 + 0.52 + 0.57 + 0.61 + 0.65 + 0.70 + 0.74 + 0.78 + 0.83 + 0.87 + 0.91 + 0.96 + 1.00) / 23) = 0.52$ . The results for all ranked average recall measurements are presented in Table 4.6 for both indexing methods.

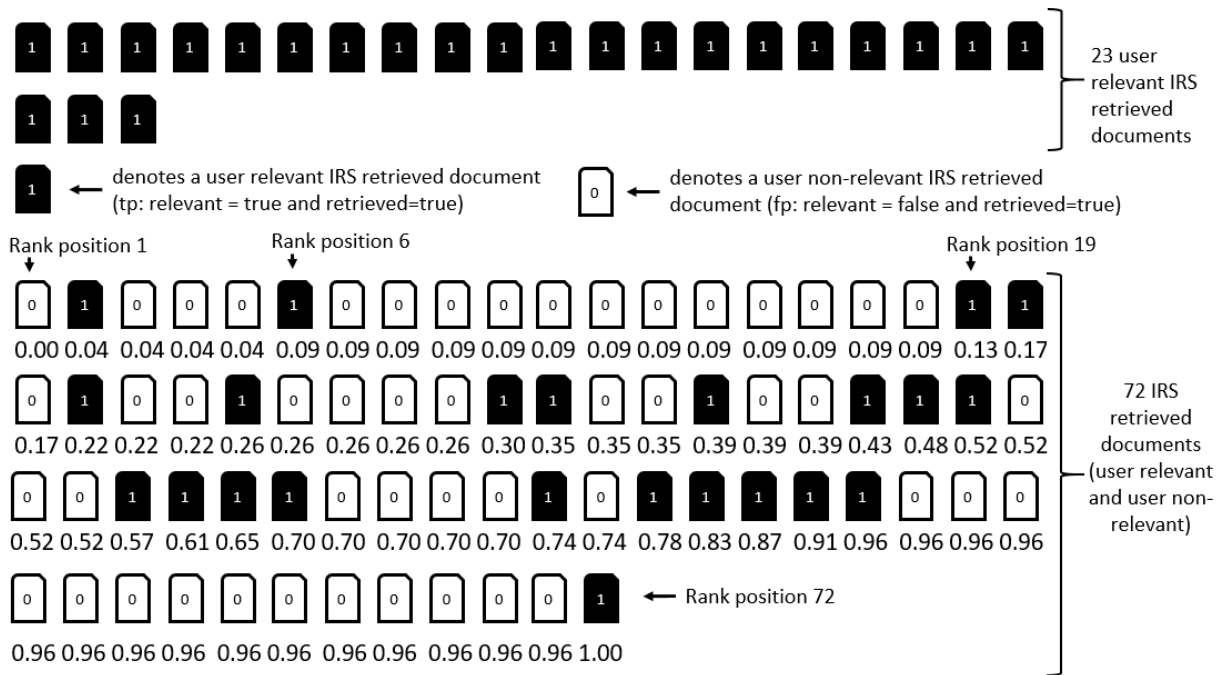


Figure 4.17: IRS-I query 1 – Recall values for one ranking of 23 relevant documents

#### 4.11.2.2 IRS-H Ranking

In the second example, the precision ranking method for IRS-H query  $q_{01}$  is explained. Out of a collection of 100 documents, query  $q_{01}$  retrieved 22 documents of which 13 were judged relevant by the user. These documents are presented in Figure 4.18.

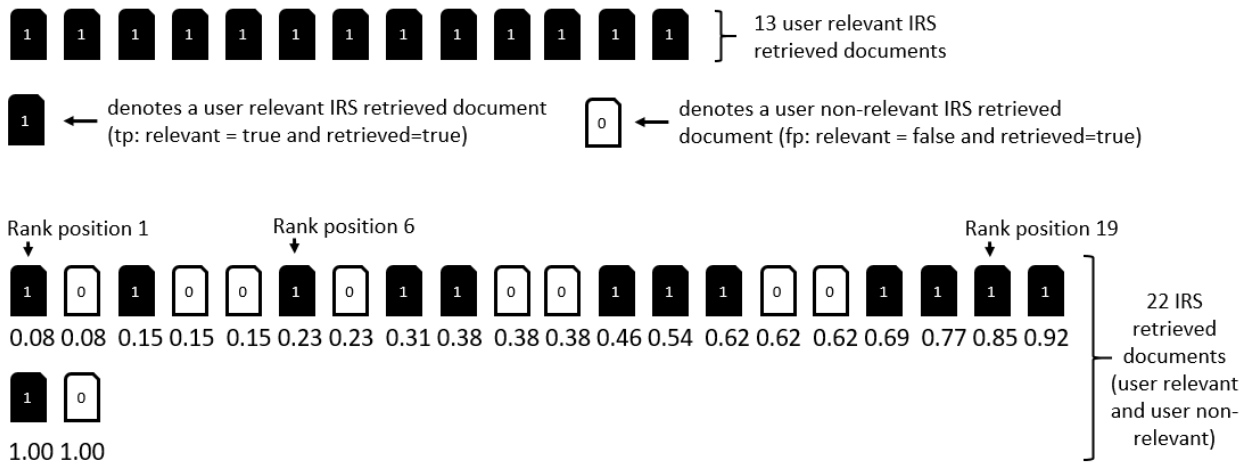


Figure 4.18: IRS-H query 1 – Recall values for one ranking of 13 relevant documents

At rank position 1, the recall value is the number of relevant documents divided by the number of user relevant retrieved documents. In this case, recall at rank position 1 is  $1 / 13 = 0.08$ , at rank position 2 is  $(1 + 0) / 13 = 0.08$ , at rank position 3 is  $(1 + 0 + 1) / 13 = 0.15$ , etc. To calculate the average recall for the query, only the rankings for the relevant documents are summed and then divided by the number of user relevant

documents, for example, Average precision for IRS-H<sub>q01</sub> =  $((0.08 + 0.15 + 0.23 + 0.31 + 0.38 + 0.46 + 0.54 + 0.62 + 0.69 + 0.77 + 0.85 + 0.92 + 1.00) / 13) = 0.54$ .

#### 4.11.3 Average recall measurements

The results for all ranked average recall measurements are presented in Table 4.6 for both indexing methods.

Table 4.6: Average recall measurements per query per indexing method

Query	IRS-I Average Recall	IRS-H Average Recall	Query	IRS-I Average Recall	IRS-H Average Recall	Query	IRS-I Average Recall	IRS-H Average Recall
q01	0.52	0.54	q26	0.67	0	q51	0.63	0
q02	0.56	0.67	q27	0.75	0	q52	0.57	1
q03	0.52	0.54	q28	0	0	q53	0.58	1
q04	0.55	0	q29	0	0	q54	0.55	1
q05	0.54	1	q30	0	0	q55	0.56	0
q06	0.52	0.53	q31	0	0	q56	0.58	0
q07	0.58	0	q32	0	0	q57	0.58	0
q08	0.63	0	q33	0.55	0.63	q58	0.55	1
q09	0.55	0	q34	0.67	0.75	q59	0.63	0
q10	0.55	0.75	q35	0.55	0	q60	0.57	0.57
q11	0.54	0.57	q36	0.67	0	q61	0.57	0.6
q12	0.57	0	q37	0.54	0.6	q62	0.58	0
q13	0.58	0	q38	0.56	0.67	q63	0.63	0
q14	0.57	0	q39	0.57	0.75	q64	0.58	0.63
q15	0.58	0	q40	0.57	0	q65	0.6	1
q16	0.55	0	q41	0.57	0	q66	0.52	0.53
q17	0.57	0.75	q42	0.54	0.63	q67	0.54	0.57
q18	0.56	0.67	q43	0.54	0.63	q68	0.55	1
q19	0.57	0	q44	0.54	0.67	q69	0.56	0.67
q20	0.58	0	q45	0.54	0	q70	0.53	0
q21	0.58	0	q46	0.55	0	q71	0.52	0.57
q22	0.58	0	q47	0.57	0	q72	0.53	0.75
q23	0.75	1	q48	0	0	q73	0.52	0.54
q24	0.57	0.75	q49	0.53	0.57	q74	0.52	0.56
q25	0.63	0	q50	0.52	0.55	q75	0.54	0.58

#### 4.11.4 Mean average recall

Mean average recall (MAR) can be expressed as the sum of the average recall values for all queries divided by the number of queries as presented in Equation 4.10 (Zhao & Huang, 2016:3).

$$MAR = \frac{\sum_{q=1}^Q AveR(q)}{Q}$$

Equation 4.10: MAR (Zhao &amp; Huang, 2016:3)

The calculated MAR results for all queries for IRS-I ( $MAR_{IRS-I}$ ) and IRS-H ( $MAR_{IRS-H}$ ) are:

$$MAR_{IRS-I} = \frac{39.39}{75} = 0.5252$$

Equation 4.11: MARIRS-I

$$MAR_{IRS-H} = \frac{26.79}{75} = 0.3572$$

Equation 4.12: MARIRS-I

As  $MAR_{IRS-I} = 0.5252$  and  $MAR_{IRS-H} = 0.3572$  then  $MAR_{IRS-H} < MAR_{IRS-I}$ . This suggests the mean average recall of IRS-H is less than the mean average recall of IRS-I. To prove statistically that these results did not occur by chance, a student's t-test was performed.

#### 4.11.5 Student's t-test and t-distribution

For the second hypothesis, a one-tailed student's t-test was used to test whether the mean average recall of IRS-H ( $MAR_{IRS-H}$ ) is less than the mean average recall of IRS-I ( $MAR_{IRS-I}$ ) and whether the result was statistically significant.

To perform the t-test, a 95% confidence level was used resulting in a significance level ( $\alpha$ ) of 5%. As there were two systems with 75 queries, the sample size  $N_q = 150$  and the degrees of freedom  $df = 148$ . The critical value ( $t_{cv}$ ) for the one-tailed t-test using a significance level of 0.05 and a  $df$  of 148 was 1.66. The MAR t-distribution and results are presented in Figure 4.19.

After performing the t-test, the t-value ( $t$ ) result was -3.565. To accept the null hypothesis  $H_{20}$  the t-value must be greater than or equal to  $t_{cv} = -1.66$ . As  $t$  is not greater than or equal to  $t_{cv}$  (as  $t < t_{cv}$ ) the null hypothesis is rejected and therefore a Type I error was made as it is a rejection of the null hypothesis. The results are statistically significant as  $p < \alpha$  where  $p$  (Type I error)  $< 0.001$  and  $\alpha = 0.05$ .

As  $t < t_{cv}$ , where  $t = -3.565$  and  $t_{cv} = -1.66$  the alternative hypothesis  $H_{21}$  is accepted and therefore:

*Hybridised indexing reduces the incorrect identification of retrieved documents.*

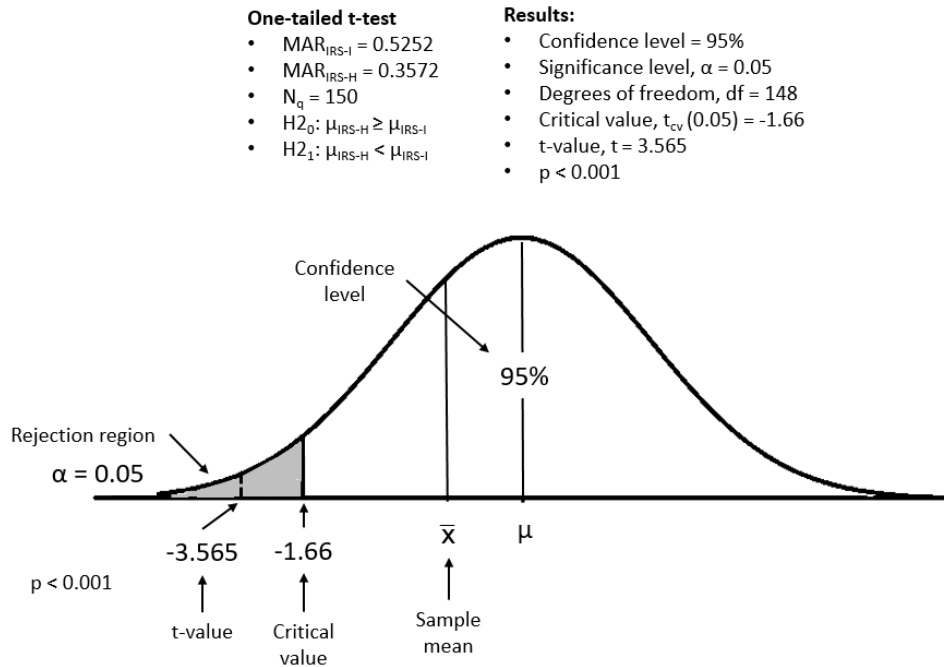


Figure 4.19: Mean average recall t-distribution and results

#### 4.12 Hypothesis 3: Analysing rejection quality of non-relevant documents

For hypothesis **H3**, the null and alternative hypotheses are stated as:

**H3<sub>0</sub>**: Hybridised indexing does not increase the quality in rejecting non-relevant documents

**H3<sub>1</sub>**: Hybridised indexing increases the quality in rejecting non-relevant documents

To test the null hypothesis for the third hypothesis, the calculations for the mean average specificity (MAS) for both indexing methods was required to determine whether or not IRS-H increases the quality in rejecting user non-relevant documents, compared to IRS-I. Thereafter, to ensure the results were not by chance, a one-tailed student's t-test was performed to verify the MAS results. The experimental results for the third hypothesis are now presented as follows: i) the specificity formula; ii) the specificity contingency table logic; iii) examples of ranking to calculate average recall per query per indexing method; iv) the presentation of the results for all ranked average specificity measurements; v) the MAS formula with examples; and vi) the

student's t-test and t-distribution results, to determine whether the results are statistically significant or not.

#### 4.12.1 Specificity measurements

Specificity can be expressed as in equation 4.13 (Cleverdon & Keen, 1966:35-36; Choudhary et al., 2017:5).

$$S = \frac{tn}{fp + tn}$$

Equation 4.13: S (Cleverdon & Keen, 1966:35-36; Choudhary et al., 2017:5)

$fp$  and  $tn$  represent user non-relevant documents (IRS retrieved and IRS not-retrieved respectively), while  $tn$  represents user non-relevant IRS not retrieved documents. The result for Specificity (refer the performance measurements in Appendix I) for IRS-I query  $q_{01}$  was calculated as follows:

$$S_{IRS-Iq_{01}} = \frac{26}{49 + 26} = 0.35$$

Equation 4.14: SIRS-Iq01

Similarly, for IRS-H, the first row represents query  $q_{01}$  and this result for Specificity was calculated as follows:

$$S_{IRS-Hq_{01}} = \frac{66}{9 + 66} = 0.88$$

Equation 4.15: SIRS-Hq01

#### 4.12.2 Ranking

One table was created for each indexing method to store the average specificity measurements per query. This resulted in 7,500 records, as there were 75 queries multiplied by 100 documents. As these tables were too large to present in this thesis, two examples are now provided of the method used to determine the average specificity measurements. The calculation of average specificity (AS) was based on the specificity values from ranked positions where relevant and non-relevant documents were retrieved for each query. If no documents were retrieved by the IRS for a query, the query's AS was set to zero.

#### 4.12.2.1 IRS-I Ranking

In the first example, the specificity ranking method for IRS-I query  $q_{01}$  is explained. Out of a collection of 100 documents, query  $q_{01}$  returned 75 documents of which 26 were rejected (these were correctly not retrieved) by IRS-I judged non-relevant by the user. These documents are presented in Figure 4.20.

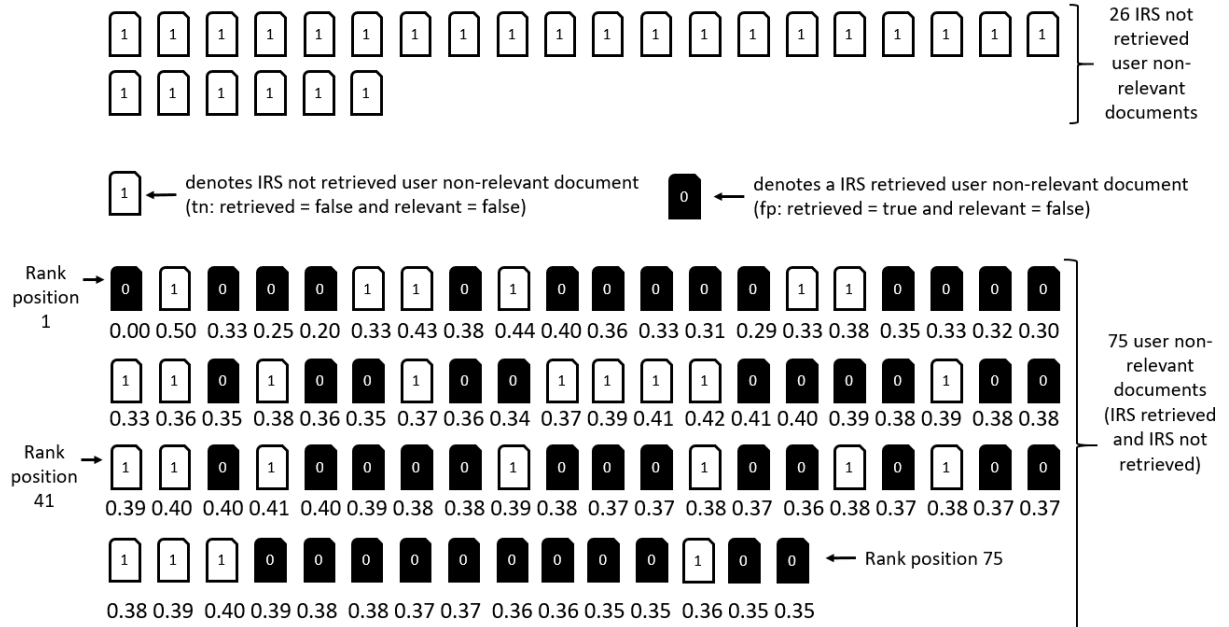


Figure 4.20: IRS-I query 1 – Specificity values of 26 rejected non-relevant documents

At rank position 1, the Specificity value is the number of not retrieved documents divided by the number of IRS not retrieved user non-relevant documents. In this case, precision at rank position 1 is  $0 / 1 = 0$ , at rank position 2 is  $(0 + 1) / (1 + 1) = 0.50$ , at rank position 3 is  $(0 + 1 + 0) / (1 + 1 + 1) = 0.33$ , etc. To calculate the average specificity for the query, only the rankings for the relevant documents are summed and then divided by the number of user relevant documents, for example, Average specificity for IRS-I $_{q_{01}}$  =  $((0.50 + 0.33 + 0.43 + 0.44 + 0.33 + 0.38 + 0.33 + 0.36 + 0.38 + 0.37 + 0.37 + 0.39 + 0.41 + 0.42 + 0.39 + 0.39 + 0.40 + 0.41 + 0.39 + 0.38 + 0.38 + 0.38 + 0.38 + 0.39 + 0.40 + 0.36) / 26) = 0.39$ . The results for all ranked average specificity measurements are presented in Table 4.7 for both indexing methods.

#### 4.12.2.2 IRS-H Ranking

In the second example, the specificity ranking method for IRS-H query  $q_{01}$  is explained. Out of a collection of 100 documents, query  $q_{01}$  returned 75 user non-relevant documents of which 9 were retrieved by the IRS. These documents are presented in Figure 4.21.

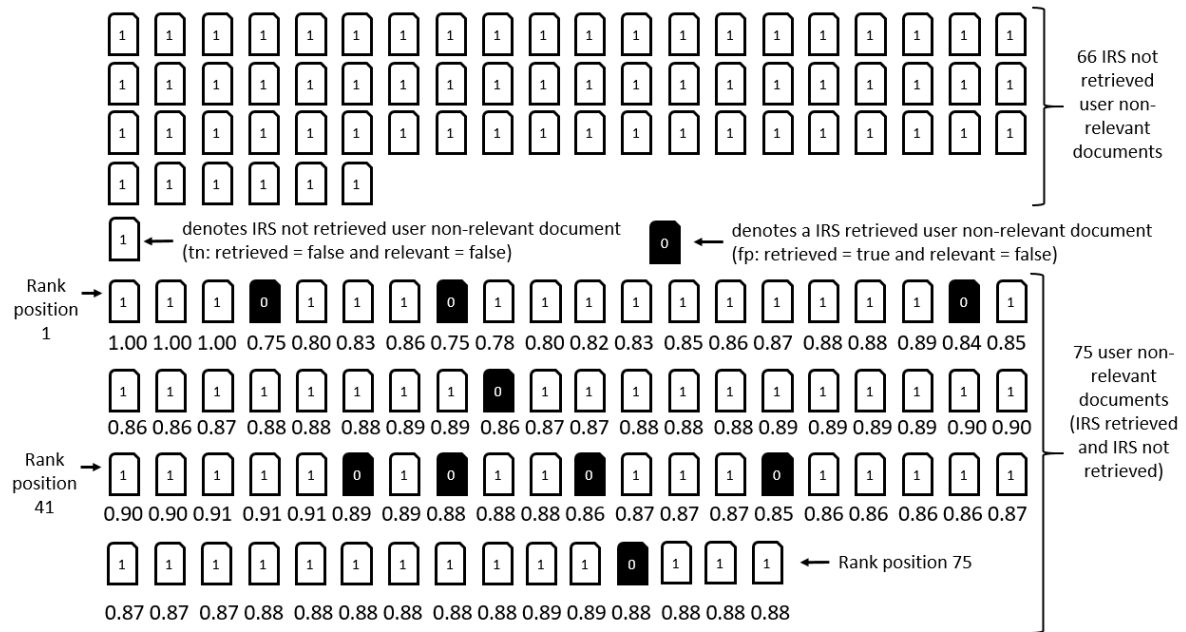


Figure 4.21: IRS-H query 1 – Specificity values of 66 rejected non-relevant documents

At rank position 1, the specificity value is the number of not retrieved documents divided by the number of IRS not retrieved user non-relevant documents. In this case, precision at rank position 1 is  $1 / 1 = 1$ , at rank position 2 is  $(1 + 1) / (1 + 1) = 1$ , at rank position 3 is  $(1 + 1 + 1) / (1 + 1 + 1) = 1$ , at rank position 4 is  $(1 + 1 + 1 + 0) / (1 + 1 + 1 + 1) = 0.75$ , etc. To calculate the average specificity for the query, only the rankings for the relevant documents are summed and then divided by the number of user relevant documents, for example, Average specificity for IRS- $l_{q01}$  =  $((1 + 1 + 1 + 0.8 + 0.83 + 0.86 + 0.78 + 0.8 + 0.82 + 0.83 + 0.85 + 0.86 + 0.87 + 0.88 + 0.88 + 0.89 + 0.85 + 0.86 + 0.86 + 0.87 + 0.88 + 0.88 + 0.88 + 0.89 + 0.89 + 0.87 + 0.87 + 0.88 + 0.88 + 0.88 + 0.89 + 0.89 + 0.89 + 0.89 + 0.9 + 0.9 + 0.9 + 0.9 + 0.91 + 0.91 + 0.91 + 0.89 + 0.88 + 0.88 + 0.87 + 0.87 + 0.87 + 0.86 + 0.86 + 0.86 + 0.86 + 0.87 + 0.87 + 0.87 + 0.87 + 0.88 + 0.88 + 0.88 + 0.88 + 0.88 + 0.88 + 0.88 + 0.88 + 0.89 + 0.89 + 0.88 + 0.88 + 0.88) / 66) = 0.87$ .

#### 4.12.3 Average specificity measurements

The results for all ranked average specificity measurements are presented in Table 4.7 for both indexing methods.

Table 4.7: Average specificity measurements per query per indexing method

Query	IRS-I Average Specificity	IRS-H Average Specificity	Query	IRS-I Average Specificity	IRS-H Average Specificity	Query	IRS-I Average Specificity	IRS-H Average Specificity
q01	0.39	0.87	q26	0.64	0.98	q51	0.67	0.98
q02	0.41	0.9	q27	0.65	0.98	q52	0.67	0.98
q03	0.43	0.93	q28	0.66	0.98	q53	0.67	0.98
q04	0.43	0.95	q29	0.67	0.98	q54	0.67	0.98



Query	IRS-I Average Specificity	IRS-H Average Specificity	Query	IRS-I Average Specificity	IRS-H Average Specificity	Query	IRS-I Average Specificity	IRS-H Average Specificity
q05	0.44	0.96	q30	0.68	0.98	q55	0.67	0.98
q06	0.42	0.97	q31	0.69	0.98	q56	0.66	0.98
q07	0.42	0.96	q32	0.7	0.98	q57	0.66	0.98
q08	0.46	0.97	q33	0.7	0.98	q58	0.66	0.98
q09	0.49	0.97	q34	0.7	0.98	q59	0.66	0.98
q10	0.51	0.97	q35	0.7	0.98	q60	0.67	0.98
q11	0.51	0.97	q36	0.7	0.98	q61	0.67	0.98
q12	0.51	0.96	q37	0.7	0.98	q62	0.67	0.98
q13	0.52	0.96	q38	0.69	0.98	q63	0.67	0.98
q14	0.53	0.97	q39	0.69	0.98	q64	0.68	0.98
q15	0.53	0.97	q40	0.7	0.98	q65	0.68	0.98
q16	0.54	0.97	q41	0.7	0.98	q66	0	0.98
q17	0.55	0.97	q42	0.7	0.98	q67	0.67	0.98
q18	0.56	0.97	q43	0.7	0.98	q68	0.66	0.98
q19	0.56	0.97	q44	0.69	0.98	q69	0	0.97
q20	0.57	0.97	q45	0.68	0.98	q70	0.64	0.97
q21	0.58	0.97	q46	0.67	0.98	q71	0	0.97
q22	0.59	0.98	q47	0.67	0.98	q72	0.62	0.97
q23	0.6	0.98	q48	0.67	0.98	q73	0	0.97
q24	0.61	0.98	q49	0.67	0.98	q74	0.61	0.97
q25	0.62	0.98	q50	0.67	0.98	q75	0.6	0.97

#### 4.12.4 Mean average specificity

Mean average specificity, based on the work of Dinh and Tamine (2015), is defined as the sum of the average specificity values for all queries divided by the number of queries as presented in equation 4.16 (Choudhary et al., 2017:5).

$$MAS = \frac{\sum_{q=1}^Q AveS(q)}{Q}$$

Equation 4.16: MAS (Choudhary et al., 2017:5)

The calculated MAS results for all queries for IRS-I ( $MAS_{IRS-I}$ ) and IRS-H ( $MAS_{IRS-H}$ ) are:

$$MAS_{IRS-I} = \frac{43.7}{75} = 0.5827$$

Equation 4.17: MASIRS-I

$$MAS_{IRS-H} = \frac{72.95}{75} = 0.9727$$

Equation 4.18: MASIRS-H

As  $MAS_{IRS-I} = 0.5827$  and  $MAS_{IRS-H} = 0.9727$  then  $MAS_{IRS-H} > MAS_{IRS-I}$ . This suggests the mean average specificity of IRS-H is greater than the mean average specificity of IRS-I. To prove statistically that these results did not occur by chance, a student's t-test was performed.

#### 4.12.5 Student's t-test and t-distribution

For the third hypothesis, a one-tailed student's t-test was used to test whether the mean average specificity of IRS-H ( $MAS_{IRS-H}$ ) is greater than the mean average specificity of IRS-I ( $MAS_{IRS-I}$ ) and whether the result was statistically significant.

To perform the t-test, a 95% confidence level was used resulting in a significance level ( $\alpha$ ) of 5%. As there were two systems with 75 queries, the sample size  $N_q = 150$  and the degrees of freedom  $df = 148$ . The critical value ( $t_{cv}$ ) for the one-tailed t-test using a significance level of 0.05 and a  $df$  of 148 was 1.66. The MAS t-distribution and results are presented in Figure 4.22.

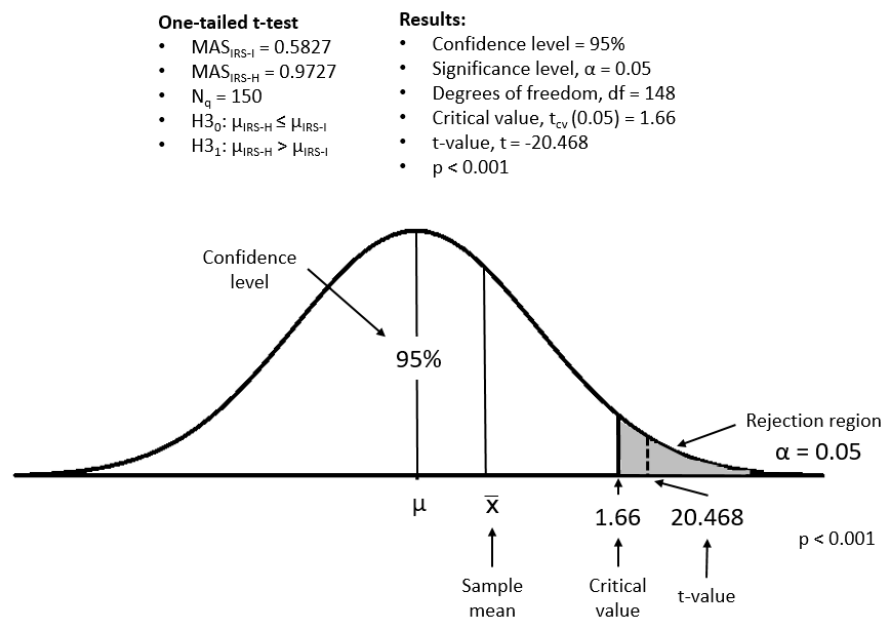


Figure 4.22: Mean average specificity t-distribution and results

After performing the t-test, the t-value ( $t$ ) result was 20.468. To accept the null hypothesis, the t-value must be less than or equal to  $t_{cv} = 1.66$ . As  $t$  is not less than or equal to  $t_{cv}$  (as  $t > t_{cv}$ ) the null hypothesis is rejected and therefore a Type I error was made as it is a rejection of the null hypothesis. The results are statistically significant as  $p < \alpha$  where  $p$  (Type I error)  $< 0.001$  and  $\alpha = 0.05$ .

As  $t > t_{cv}$ , where  $t = 20.468$  and  $t_{cv} = 1.66$  the alternative hypothesis **H3<sub>1</sub>** is accepted and therefore:

*Hybridised indexing increases the quality in rejecting non-relevant documents.*

At this stage of the research with the results presented, the research question has been answered, and with the three hypotheses **H1**, **H2** and **H3** having been tested, the three alternative hypotheses were accepted. Thus, these three hypotheses statistically proved that the hybrid indexing method worked. In addition, and in doing so, the results indicate that IRS-H performed better than IRS-I since:

- i) IRS-H increased the effectiveness of retrieving relevant documents compared with IRS-I,
- ii) IRS-H reduced the incorrect identification of retrieved documents compared with IRS-I, and
- iii) IRS-H increased the quality in rejecting non-relevant documents compared with IRS-I.

#### **4.13 Hypothesis 4: Judgments made by IRS-H and the user**

The focus of this research now shifts to addressing the research problem: i) by analysing the judgements made between IRS-H and the user; and ii) by attempting to satisfy the information needs of the user. This is performed by testing the two hypotheses **H4** and **H5** quantitatively using the Kappa coefficient and a range of agreement measurements to test the strength of agreement between the judgments made by IRS-H and those of the users. The testing of hypotheses **H4** and **H5** follow the IRS-H versus user system and user-generated judgment flow chart, described in section 3.5.2.4, Figure 3.14.

For hypothesis **H4**, the null and alternative hypotheses are stated as:

**H<sub>40</sub>**: Judgments made by the hybrid indexing method and the user disagree

**H<sub>41</sub>**: Judgments made by the hybrid indexing method and the user agree

The judgement results generated by IRS-H and by the users from the questionnaire are now presented. The results from the questionnaires completed by the five users (the user) were answers to the questions. At the query level, these results were based on the Boolean true or false values which were then converted to binary, where 1 represented true and 0 represented false. For IRS-H, the results were stored in a phrase-term-by-document matrix. Within this matrix, the columns represented the phrase-terms, the rows represented the documents, and each cell stored the phrase-term frequency (*ptf*), the number of times a phrase-term occurred in a document. This data were summarised to the query level and converted to binary where 1 represented  $ptf > 0$  and 0 represented  $ptf = 0$ . In summary, the system judgements were produced by IRS-H and the results from the questionnaire were produced by the user judgements. These IRS-H and user judgements were made for each of the 65 phrase-

term queries for each of the 100 documents in the collection. In total, IRS-H and the users made 6,500 judgements, 1,300 per user. To test whether the vocabulary mismatch problem could be solved, there was a need to compare the system-generated results from IRS-H with those of the questionnaire-generated results from the users.

To test the strength of agreement between these judgments, Kappa coefficient tests based on the work of Cohen (1960) were performed comparing the system-generated results from IRS-H with those of the five users. The Kappa coefficient measures the consistency of agreement between two judgements and the suggested range for these measurements was taken from the work of Landis and Koch (1977) and Fleiss et al. (2003). This range has six divisions from 'poor' to 'almost perfect', indicated by values ranging from less than 0 to 1. This six-division range for the Kappa coefficient is listed in Table 4.8.

**Table 4.8: Kappa coefficient agreement measures (Landis & Koch, 1977:7)**

Kappa coefficient	Strength of agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

To test whether judgments made by IRS-H and the user agreed, the judgments between IRS-H and the users were measured and the results presented in Table 4.9. When completing the questionnaire during the experiment, the users were requested to indicate whether a specific phrase-term existed in a document. The aim was for each user to read through the document and if the phrase-term existed in the document, the users then indicated this by placing a tick against the phrase-term. Below are the results per user: each answered 65 questions pertaining to the 65 phrase-terms from a collection of 20 documents, totalling 1,300 cases. The judgement results made by IRS-H were then compared with each of the five users' results. The Kappa coefficient ( $k$ ) and significance ( $p$ ) were then calculated using SPSS. To determine the matching of judgements and the strength of agreement between the judgments made by the IRS-H and the five users, the six-division range proposed by Landis and Koch (1977) and Fleiss et al. (2003) was used. These results are presented in Table 4.9.

Table 4.9: IRS-H versus user judgements – phrase-terms

Judgement test	Query structure	No of docs	No of cases	Statistical results					
				User	Kappa coefficient ( $k$ )	Strength of agreement	Significance ( $p$ )	Statistically significant	Rank
IRS-H + User	65 single queries each containing a single phrase-term	20	1300	E	0.672	Substantial	$p < 0.001$	Yes	1
				C	0.491	Moderate	$p < 0.001$	Yes	2
				B	0.487	Moderate	$p < 0.001$	Yes	3
				D	0.385	Fair	$p < 0.001$	Yes	4
				A	0.032	Slight	$p < 0.001$	Yes	5

From the statistically significant results, the user with the highest rank, with a substantial strength of agreement with IRS-H – the second best according to Landis and Koch (1977) and Fleiss et al. (2003) – was User E where  $k_E = 0.672$  and  $p_E < 0.001$ . Ranked second and third were users C and B with a moderate strength of agreement with  $k_C = 0.491$  and  $p_C < 0.001$ , and  $k_B = 0.487$  and  $p_B < 0.001$  respectively. User D dropped to a fair strength of agreement with  $k_D = 0.385$  and  $p_D < 0.001$ . Ranked fifth, User A had the least matching judgments with a slight strength of agreement with  $k_A = 0.032$  and  $p_A < 0.001$ . The results show that a mismatch between the judgements of the user and IRS-H still exists and that the mismatch between users' judgements differs widely.

To determine why the user judgements differed widely, the phrase-term-by-document matrix was reviewed and a physical investigation was made by this researcher to verify whether a phrase-term occurred at least once in a document. Further analysis was performed: firstly, a physical check was performed by this researcher for each phrase-term for each document. The judgements matched perfectly for all cases except for four. Four documents contained phrase-terms that IRS-H matched where commas or brackets were situated in-between. As these special characters were removed during the information gathering process, these were erroneously matched and retrieved by IRS-H. The two-word phrase-term  $pt_{01}$  'design science' was matched incorrectly to 'design, science' in the text three times in documents  $d_{0009}$ ,  $d_{0010}$  and  $d_{0060}$ , and the two-word phrase-term  $pt_{09}$  'qualitative method' was matched incorrectly to 'qualitative) method' in the text once in document  $d_{0048}$ .

As the results show that a mismatch between the judgements of the user and IRS-I, and to a lesser extent between the users and IRS-H, still exists, the key findings are:

- i) There is a significant disparity between the judgements of the five users. Strength of agreements range from slight to fair, to moderate, to substantial. All users missed matching phrase-terms to documents thus creating this disparity.

- ii) All users incorrectly stated at least five times that a match did not occur between a phrase-term and its existence within a document, when in fact it did. The best judgements made in rank order were: User A, User E, User C, User B and User D with mismatching percentages below 2.5% (Table 4.10).

**Table 4.10: IRS-H versus User – mismatched judgement cases**

User	Total judgement cases	Mismatched cases User=0 IRS-H=1	Mismatch %	Rank
A	1300	5	0.38%	1
E	1300	22	1.69%	2
C	1300	26	2.00%	3
B	1300	28	2.15%	4
D	1300	31	2.38%	5

- iii) All users incorrectly stated a match occurred at least twice between a phrase-term and its existence within a document, when it did not. The best judgements made in rank order were: User C, User E, User D and User B with mismatching percentages below 1.5%. However, the mismatching percentage for User A, ranked fifth, was a very high 62.23% with 809 mismatching cases. Further analysis was performed and it was discovered that User A incorrectly stated that all phrase-terms existed when the information need was judged relevant (Table 4.11).

**Table 4.11: User versus IRS-H – mismatched judgement cases**

User	Total judgement cases	Mismatched cases User=1 IRS-H=0	Mismatch %	Rank
C	1300	2	0.15%	1
E	1300	5	0.38%	2
D	1300	11	0.84%	3
B	1300	19	1.46%	4
A	1300	809	62.23%	5

None of the users agreed with IRS-H with a perfect score of  $k = 1.00$ , and none of the users agreed with IRS-H with an almost perfect score of  $0.8 \leq k \leq 1.00$ . Referring to Table 4.9,  $k \neq 1.00$  for any of the five users and therefore the null hypothesis  $H_{4_0}$  is accepted:

*Judgments made by the hybrid indexing method and the user disagree.*

#### 4.14 Hypothesis 5: Satisfying the user's information need

For hypothesis **H5**, the null and alternative hypotheses are stated as:

**H5<sub>0</sub>**: The hybrid indexing method does not satisfy the information needs of the user

**H5<sub>1</sub>**: The hybrid indexing method satisfies the information needs of the user

To test the hypothesis, the judgements made by each of the five users were compared with those judgements made by IRS-H. When completing the questionnaire during the experiment, the users were requested to indicate whether a document was relevant to a specific information need. The aim was for each user to read through the document and if the document was judged relevant by the user, this was indicated by using a tick against the information need pertaining to that document. Table 4.12 presents the results per user: each answered ten questions pertaining to the ten information needs from a collection of 20 documents, totalling 200 cases. The judgement results made by IRS-H were then compared with each of the five users' results. The Kappa coefficient ( $k$ ) and significance ( $p$ ) were then calculated using SPSS. To determine the strength of agreement between the judgments made by the IRS-H and the five users, the six-division range proposed by Landis and Koch (1977) and Fleiss et al. (2003), was utilised.

Table 4.12: IRS-H versus User judgements – information needs

Judgement test	Query structure	No of docs	No of cases	Statistical results					
				User	Kappa coefficient ( $k$ )	Strength of agreement	Significance ( $p$ )	Statistically significant	Rank
IRS-H * User	10 expanded queries, one per information need	20	200	B	0.502	Moderate	$p < 0.001$	Yes	1
				E	0.400	Fair	$p < 0.001$	Yes	2
				C	0.290	Fair	$p < 0.001$	Yes	3
				D	0.204	Slight	$p < 0.001$	Yes	4
				A	0.153	Slight	$p < 0.001$	Yes	5

From the statically significant results the user with the highest rank, with a moderate strength of agreement with IRS-H, was User B where  $k_B = 0.502$  and  $p_B < 0.001$ . Ranked second and third were users E and C with a fair strength of agreement with  $k_E = 0.400$  and  $p_E < 0.001$ , and  $k_C = 0.290$  and  $p_C < 0.001$  respectively. Users D and A dropped to a slight strength of agreement with  $k_D = 0.204$  and  $p_D < 0.001$ , and  $k_A = 0.153$  and  $p_A < 0.001$ , respectively. From the results, it is evident that IRS-H does not satisfy the information need of the user.

Since the user results differ from those of IRS-H, it is evident that the users are dissatisfied with the information needs as judged by IRS-H.

Referring to Table 4.12,  $k \neq 1.00$  for any of the five users and therefore the null hypothesis  $H5_0$  is accepted:

*The hybrid indexing method does not satisfy the information needs of the user.*

Referring to Table 4.12, as none of the users agreed with IRS-H with a perfect score of  $k = 1.00$ , and none of the users agreed with IRS-H with an almost perfect score of  $0.8 \leq k \leq 1.00$ , none of the users agreed that IRS-H satisfied their information needs. Although the hybridised indexing method succeeded in exact matching of a query to a document, and succeeded in maintaining word ordinality and proximity, the effect from the users' results indicate that the hybrid indexing method did not satisfy the information needs of the user. Clearly there is much more work still to be performed within this area of research.

#### 4.15 Summary of experimental findings

The summary of the findings for the experiment comparing users' judgements to those of IRS-H using the hybrid indexing method, and users' judgements to those of IRS-I using the inverted indexing method, are presented in Table 4.13.

**Table 4.13: Summary of experimental findings**

No	Finding
1	IRS-H performed better than IRS-I as the hybridised indexing method increased the effectiveness of retrieving relevant documents.
2	IRS-H performed better than IRS-I as the hybridised indexing method reduced the incorrect identification of retrieved documents.
3	IRS-H performed better than IRS-I as the hybridised indexing method increased the quality in rejecting non-relevant documents.
4	There is a significant disparity between the judgements of the five users. Strength of agreements range from slight to fair, to moderate, to substantial. All users missed matching phrase-terms to documents thus creating this disparity.
5	All users incorrectly stated at least five times that a match did not occur between a phrase-term and its existence within a document when in fact it did. The best judgements made in rank order were: User A, User E, User C, User B and User D with mismatching percentages below 2.5%.
6	All users incorrectly stated a match occurred at least twice between a phrase-term and its existence within a document when it did not. The best judgements made in rank order were: User C, User E, User D and User B with mismatching percentages below 1.5%. However, the mismatching percentage for User A, ranked fifth, was a very high 62.23% with 809 mismatching cases. Further analysis was performed and it was discovered that User A incorrectly stated all phrase-terms existed when the information need was judged relevant.
7	None of the users agreed with IRS-H with a perfect score of $k = 1.00$ , and none of the users agreed with IRS-H with an almost perfect score of $0.8 \leq k \leq 1.00$ . $k \neq 1.00$ for any of the five users and therefore judgments made by the hybrid indexing method and the user disagree.



No	Finding
8	As none of the users agreed with IRS-H with a perfect score of $k = 1.00$ , and none of the users agreed with IRS-H with an almost perfect score of $0.8 \leq k \leq 1.00$ , none of the users agreed that IRS-H satisfied their information needs. Although the hybridised indexing method succeeded in exact matching of a query to a document, and succeeded in maintaining word ordinality and proximity, the effect from the users' results indicate that the hybrid indexing method did not satisfy the information needs of the user. Clearly there is much more work still to be performed within this area of research.

#### 4.16 The second research question

To complete Section B, the second research question is now addressed. The research questions is:

##### **RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

In all three pilots and during the evaluation, exact matches occurred for all phrase-terms expressed in queries to those phrase-terms that existed within the documents.

To answer the second research question:

*The hybrid index design solves the vocabulary mismatch problem of matching a query to a document.*

#### 4.17 Summary

A summary of the results for this chapter is now presented:

- i) For the first research question: **How can an IRS index be designed that maintains word ordinality and word proximity?**

The design of the hybrid indexing method is one way an IRS can be designed to maintain word ordinality and word proximity. The hybrid indexing method utilises a pair of hybrid indices: the hybrid token index and the hybrid query index. During the information gathering process, the hybrid token index of IRS-H is populated with tokens acquired from the text and each is allocated a unique Token ID. This Token ID is the key concept in this design and maintains the ordinality of the words and the proximity of words. When the search engine process is activated, the words within the query's multi-word phrase-terms are matched exactly to the tokens in the hybrid token index maintaining word ordinality and proximity.

- ii) For the three hypotheses: IRS-H versus IRS-I:
- *The alternative hypothesis  $H1_1$  is accepted: Hybridised indexing increases the effectiveness of retrieving relevant documents.*
  - *The alternative hypothesis  $H2_1$  is accepted: Hybridised indexing reduces the incorrect identification of retrieved documents.*
  - *The alternative hypothesis  $H3_1$  is accepted: Hybridised indexing increases the quality in rejecting non-relevant documents.*
- iii) For the two hypotheses: IRS-H versus user:
- *Based on the results the null hypothesis  $H4_0$  is accepted: Judgments made by the hybrid indexing method and the user disagree*
  - *Based on the results the null hypothesis  $H5_0$  is accepted: the hybrid indexing method does not satisfy the information needs of the user.*
- iv) For the second research question: **Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

During Pilot testing and evaluation, the hybrid index design solved the vocabulary mismatch problem between matching a query to a document.

This chapter has taken the exploratory and design science research approach to present a novel design of a hybrid indexing method used within IRS-H. The results from this design helped answer the first research question. In addition, this chapter has taken the explanatory and experimental approach to generate system data from IRS-H and IRS-I, and user-generated data using questionnaires, to enable the five hypotheses to be tested and to answer the second research question.

## CHAPTER FIVE: DISCUSSION

*"Knowing is not enough; we must apply. Willing is not enough; we must do"*  
 – Johann Wolfgang von Goethe

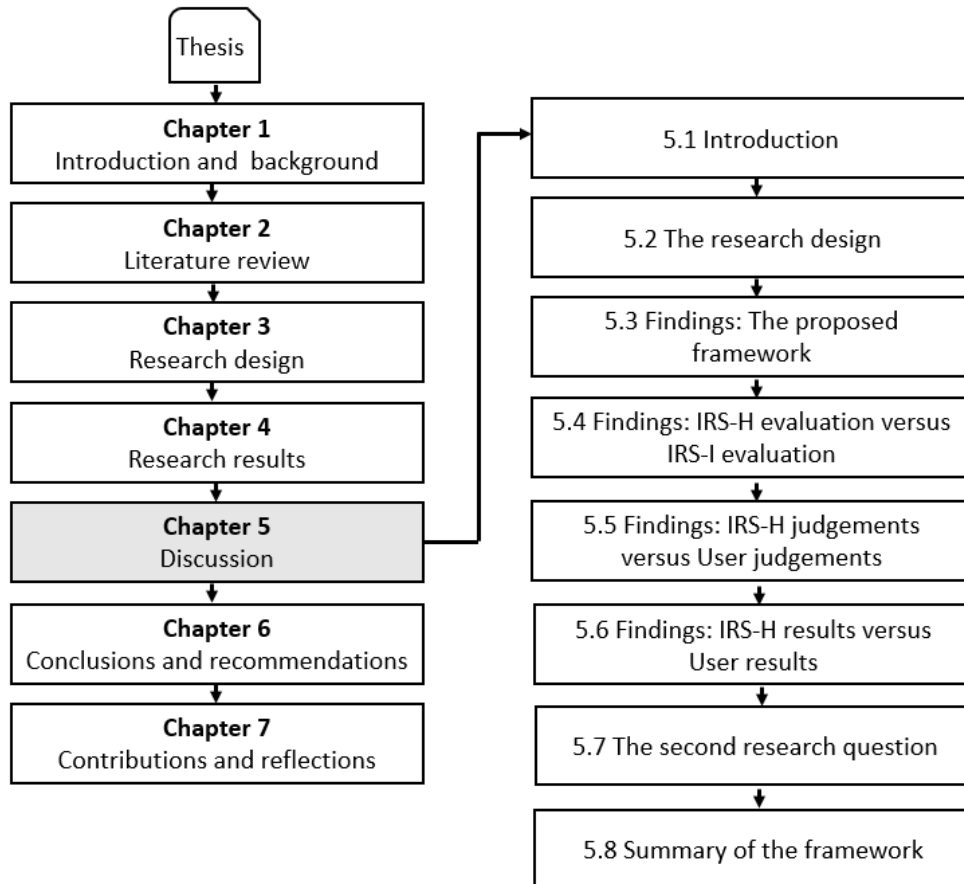


Figure 5.1: Schematic representation of Chapter Five

### 5.1 Introduction

Chapter Five revisits the research design and the proposed framework, and summarises the key findings. As the research design is complex, an illustration of how this discussion chapter is presented is provided in Figure 5.2. Reading from left-to-right there are six columns, the first five columns relate to the research design including: i) the research problem; ii) the aim of the research; iii) the objectives of the research; iv) the hypotheses; and v) the research questions. The sixth column relates to the four main sections provided in this discussion chapter (sections 5.3, 5.4, 5.5 and 5.6). For the benefit of the reader, section 5.2.1 re-states the research problem followed by section 5.2.2 that re-states the two aims of this research. The six objectives required for the two research questions are then presented in section 5.2.3 (Note, following the flow in Figure 5.2, the first objective had to be met before the second aim and its six objectives could proceed). Column 4 reiterates the five hypotheses (section 5.2.4) and column 5 states the two research questions (sections

5.2.5 and 5.7). The sixth column describes the final four sections where a discussion of the results takes place for: i) the conceptual framework of the design of the hybrid indexing method (section 5.3); ii) the key findings determined when comparing the effectiveness of IRS-H with IRS-I (section 5.4); iii) the key findings determined when comparing judgements made by IRS-H with the user (section 5.5); and iv) the key findings determined when comparing the overall results generated by IRS-H and the user. The findings from the first research question, the hypotheses, IRS evaluations, IRS judgements, and IRS results are collated, and the second research question (section 5.7) is then presented followed by a discussion.

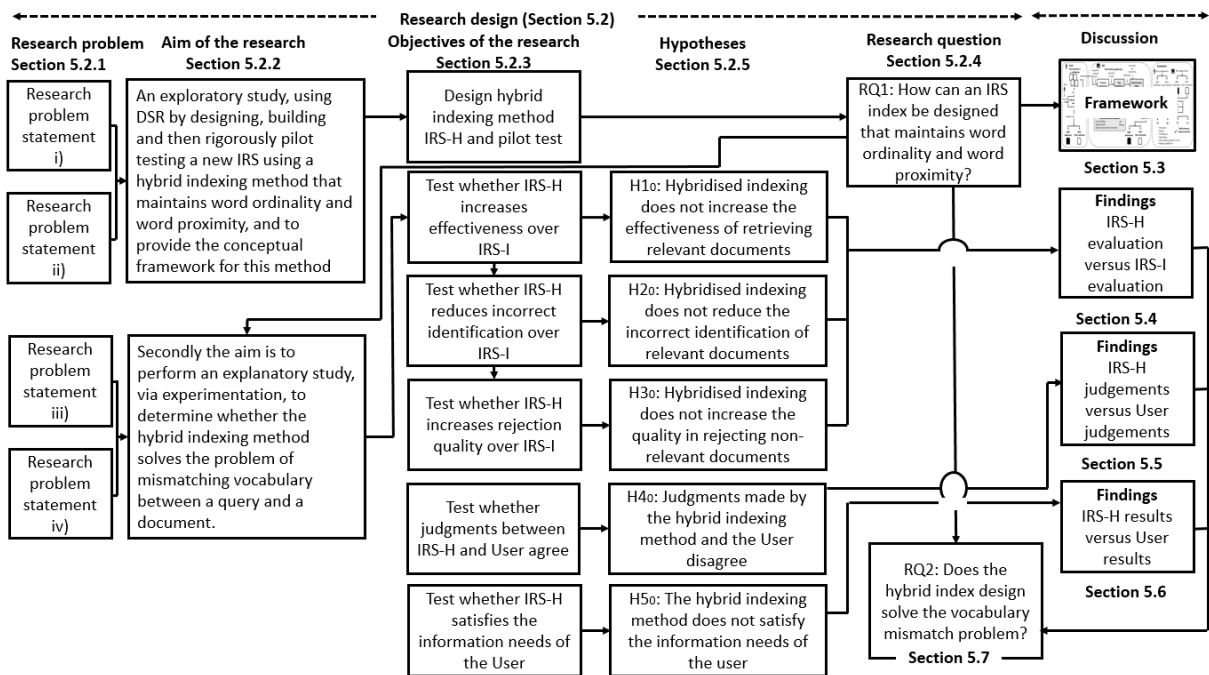


Figure 5.2: The layout of Chapter Five

## 5.2 The research design

This section revisits the research problem, the aims of the research, the objectives of the research, the research questions and hypotheses, and there is discussion to confirm that these requirements have been met. A discussion of the key findings is presented in later sections.

### 5.2.1 The research problem

In section 1.3, the research problem for this study was stated as:

Challenges exist for information retrieval systems in handling mismatching vocabularies in queries and candidate source documents (Onal et al., 2018). As a result, these information retrieval systems may retrieve some documents that are non-relevant and miss some that are relevant (Van Gysel, 2017). This increases the time

for research by forcing additional perusal of unsatisfactory results, and additional searches using alternative vocabularies (Liu et al., 2017). This renders information retrieval systems less effective than they could be, and inhibits productive research (Mitra & Awekar, 2017; Nguyen et al., 2018).

Referring to Figure 5.3, the research problem consists of four distinct statements, which feed into two aims. The four statements are:

- i) Challenges exist for information retrieval systems in handling mismatching vocabularies in queries and candidate source documents (Onal et al., 2018).
- ii) As a result, these information retrieval systems may retrieve some documents that are non-relevant and miss some that are relevant (Van Gysel, 2017).
- iii) This increases the time for research by forcing additional perusal through unsatisfactory results, and additional searches using alternative vocabularies (Liu et al., 2017).
- iv) This renders information retrieval systems less effective than they could be, and inhibits productive research (Mitra & Awekar, 2017; Nguyen et al., 2018).

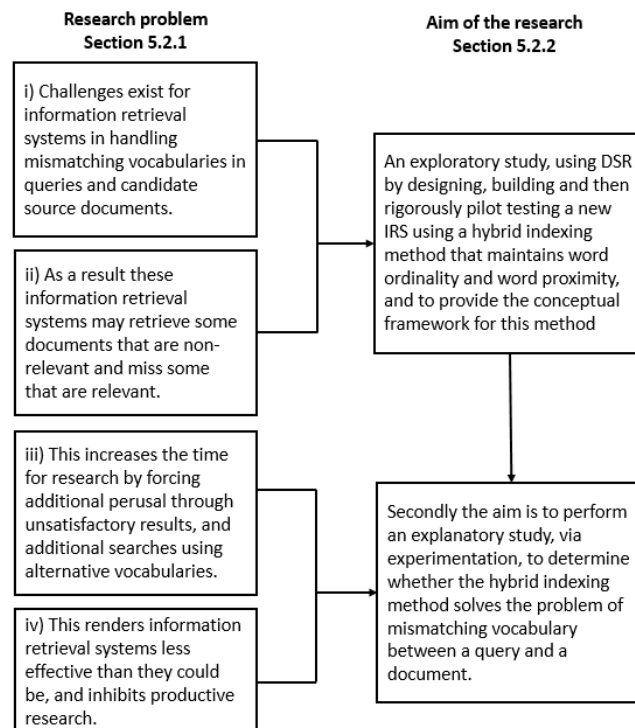


Figure 5.3: The research problem and aims

In order to address the first and second statements of the research problem, the aim was to perform an exploratory study (Robson, 2005) using DSR (Hevner et al., 2019) to design and build a new method of hybridised indexing that could:

- i) solve the many challenges in mismatching vocabularies between queries and documents,

- ii) reduce the non-relevant documents retrieved, and
- iii) prevent missing the relevant documents through exact matching of queries to documents (section 5.2.2).

Once the new hybridised indexing method was developed and had been shown to work through rigorous pilot testing, the third and fourth statements of the research problem were addressed in order to be resolved. At this stage, the aim was to perform an explanatory study (Babbie, 2013; Saunders et al., 2019) using experimentation, analytics and the quantitative method, by being more specific, in order to try to decrease these unsatisfactory results, by improving the quality in the rejection of non-relevant documents and by increasing overall system effectiveness (section 5.2.2).

### **5.2.2 The aim of the research**

In order to investigate the research problem, two aims were stated (section 1.5):

- i) The first aim was to conduct an exploratory study, using DSR by designing, building and then rigorously pilot testing a new IRS using a hybrid indexing method that maintained word ordinality and word proximity, and to provide the conceptual framework for this method.

To assist in achieving the first aim, one objective was set (section 5.2.3).

- ii) The second aim was to perform an explanatory study, via experimentation, to determine whether the hybrid indexing method solved the problem of mismatched vocabulary between a query and a document.

To assist in achieving the second aim, six objectives were set (section 5.2.3).

### **5.2.3 The objectives of the research**

Following the two aims of this research, seven objectives were set, one relating to the exploratory study and the remaining six relating to the explanatory study. For the exploratory study the single aim was: to design, build, and rigorously pilot test a hybrid indexing method that maintains word ordinality and word proximity (IRS-H), and to compare the effectiveness of this method with traditional inverted indexing method (IRS-I).

For the explanatory study the six objectives were:

- i) to test whether an IRS using a hybrid indexing method increases the effectiveness of retrieving only those documents that are judged relevant by the user,
- ii) to test whether the hybrid indexing method reduces errors in incorrect identification of user judged relevant documents, thus reducing the number of documents for the user to peruse,
- iii) to test whether the hybrid indexing method increases the rejection quality of user non-relevant documents, thus providing confidence to the user in the judgement of the IRS,
- iv) to determine whether the judgments made by the hybrid indexing method and the user agree,
- v) to determine whether the hybrid indexing method satisfies the information needs of the user by retrieving those documents from the collection that are relevant to the user, and
- vi) to determine whether the hybrid indexing method solves the problem of mismatching vocabulary between a query and a document.

#### **5.2.4 The first research question**

To achieve the first objective, the first research question was set and had to be answered before the subsequent six objectives could be addressed and before the second research question could be answered. The first research question was thus:

**RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

Sections 4.3, 4.4, 4.5 and 4.6 explain in detail how the pair of hybrid indices have been designed and built and through rigorous pilot testing (Appendices A, B and C) have been shown to work. By answering the first research question, the first objective was met (section 4.7). Once the first research question had been answered and it had been shown that the hybrid indexing worked and could maintain word ordinality and word proximity, through pilot testing, and thus could match a query to a document exactly, then following this, the three evaluation hypotheses were tested.

#### **5.2.5 The hypotheses**

##### **5.2.5.1 The evaluation hypotheses**

The three null hypotheses related to the evaluation of IRS-H with that of IRS-I were:

**H1<sub>0</sub>:** Hybridised indexing does not increase the effectiveness of retrieving relevant documents

**H2o:** Hybridised indexing does not reduce the incorrect identification of relevant documents

**H3o:** Hybridised indexing does not increase the quality in rejecting non-relevant documents

Each of these three null hypotheses was rejected and the alternative hypotheses accepted (sections 4.10, 4.11 and 4.12). By performing these tests on these three hypotheses, objectives 2, 3 and 4 were met.

#### **5.2.5.2 The judgment hypothesis**

The single null hypothesis related to the judgements between IRS-H and that of the user was:

**H4o:** Judgments made by the hybrid indexing method and the user disagree

The null hypothesis was accepted (section 4.13). By performing this test on this single hypothesis, objective 5 (test whether judgments between IRS-H and user agree) was met.

#### **5.2.5.3 The results hypothesis**

The single null hypothesis related to the results generated between IRS-H and that of the user was:

**H5o:** The hybrid indexing method does not satisfy the information needs of the user

The null hypothesis was accepted (section 4.14). By performing this test on this single hypothesis, objective 6 (test whether IRS-H satisfies the information needs of the user) was met.

#### **5.2.6 The second research question**

To answer the second research question, the first six objectives had to be met and the first research question answered, with a definitive indexing design that could maintain word ordinality and proximity. In addition, the testing of the first five hypotheses had to have been concluded.

These results provided input to assist in answering the second research question:



**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

To answer this research question, confirmation was required that IRS-H could match a query to a document. The design allowed for a perfect match, and during pilot testing (Appendices A, B and C) these matches were confirmed to be accurate. Sections 5.5 and 5.6 explain the key findings and section 5.7 answers the second research question in further detail.

**5.3 Findings: the proposed framework**

The aim of the first research question was to design, build, and rigorously pilot test a hybrid indexing method that maintains word ordinality and word proximity (IRS-H), and to compare the effectiveness of this method with the traditional inverted indexing method (IRS-I). To fulfil this research objective, the design of this IRS index, which is a pair of indices that uses a hybrid token index and a hybrid query index, is explained in Chapter Four. The physical design of the indices and their relationships was explained using ERDs. Examples of populated tables were provided. For a complete view of the design and build of these indices, refer to the three pilot tests based on DSR (Hevner, 2007) in Appendices A, B and C. The results were promising as the indexing method maintained word ordinality and word proximity for phrase-terms that contained one, two, three, four, or more words. In the data analysis section of Chapter Four, a number of statistics were produced. These statistics were used to test the first three hypotheses and to prove that the hybrid indexing method worked.

Looking back on the theoretical conceptual framework presented in Chapter Two, this first framework was based upon the inverted indexing method (Croft et al., 2015) using theory discussed during the literature review. However, from lessons learned during this research process, this framework evolved, as specific design changes addressed the need to achieve effective results.

After a thorough review of the literature, including the theories, methods and concepts of information retrieval, and taking cognisance of all five hypotheses for this research, the definitive framework of the hybrid indexing method for this study is now presented in Figure 5.4.

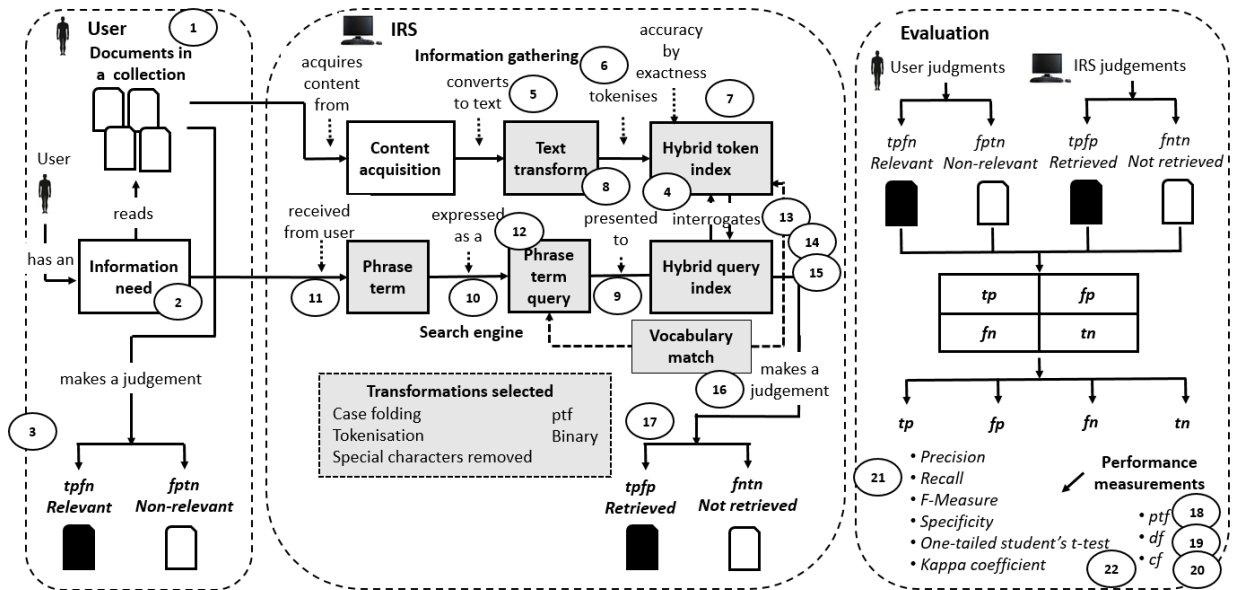


Figure 5.4: The proposed framework of the hybrid indexing method

There are three stages to the process where the hybrid indexing method is used: the user stage, the IRS stage, and the evaluation stage. Referring to Figure 5.2, column six, the first key finding is the presentation of the proposed framework of the hybrid indexing method. The results show (sections 4.3, 4.4, 4.5, 4.6 and 4.7) that the hybrid indexing method is capable of exact matching single-word or multi-word phrases (Ha et al., 2002) expressed within a query to those single-word or multi-word phrases that exist within a document. From the new design, by using a pair of indices (the hybrid token index and the hybrid query index) and by using the concept of a unique token ID within these indices, word ordinality (the sequencing of words) and word proximity (the distance between words) are maintained precisely.

**Finding 1:** *The hybrid indexing method maintains word ordinality and word proximity*

**Finding 2:** *The hybrid indexing method matches a query to a document exactly*

Table 5.1 illustrates the key design findings of the method. There are five columns: i) indicates the design number; ii) indicates the stage within the hybrid indexing method (User, IRS or Evaluation stage); iii) Process identifies which process is being referred to; iv) the inverted indexing method describes the design concepts from the literature; v) the hybrid indexing method describes the updated design concepts for the method used in this research; and vi) a reference where applicable.

Table 5.1: Design findings of the hybrid indexing method

No	Stage	Process	The inverted indexing method	The hybrid indexing method	Reference
1	User	Document collection	Documents are collected to provide a closed document collection	No change – documents are collected to provide a closed document collection	Van Gysel et al. (2017)
2	User	Information need	Information needs are provided by the user	No change – information needs are provided by the user	Singhal (2001)
3	User	Relevance judgement	User judges whether a document is relevant or non-relevant	No change – User judges whether a document is relevant or non-relevant	Croft et al. (2015)
4	IRS	Information gathering	The inverted indexing method uses a single index populated with tokens	The hybrid indexing method uses a pair of indices: a token index and a query index	Van Gysel (2017)
5	IRS	Information gathering	Text remains unchanged	Text is case folded to lowercase with special characters removed	Ruthven and Lalmas (2003)
6	IRS	Information gathering	Tokens are distinct	Tokens are non-distinct but have unique token IDs – combined as a key makes them distinct	Ha et al. (2002)
7	IRS	Information gathering	The inverted index contains fewer tokens than the token index	The token index expands containing more tokens than the inverted index	N/A
8	IRS	Information gathering	Stemming, stopping, suffix stripping, vectors and other techniques may be used to improve impreciseness of the system	Stemming, stopping, suffix stripping, vectors and other techniques are unnecessary owing to the exact matching capability of the hybrid indexing method	Tordai (2006)
9	IRS	Search engine	A query reads the index and returns a result	A query populates the query index which interrogates the token index and returns a result	Manning et al. (2008)
10	IRS	Search engine	Queries contain single-word terms	Queries contain single- or multi-word phrase-terms	Clarke et al. (2000)
11	IRS	Search engine	Words expressed as terms remain unchanged	Words in phrase-terms are case folded to lowercase with special characters removed	N/A
12	IRS	Search engine	Terms are treated individually	Phrase-terms are treated individually but the words within the phrase-terms are treated as a whole	N/A
13	IRS	Search engine	A query attempts to match each term to a token in the inverted index	A query attempts to match each phrase-term within the query index to a range of tokens within the token index	N/A
14	IRS	Search engine	Word ordinality is ignored	Word ordinality is maintained	N/A
15	IRS	Search engine	Word proximity is ignored	Word proximity is maintained	N/A

No	Stage	Process	The inverted indexing method	The hybrid indexing method	Reference
16	IRS	Search engine	A query attempts to match at least one of the words in the query to a word in a document. If an approximate match occurs, the document number is returned.	A query attempts match all words in a phrase-term to a string of text in a document exactly. If an exact match occurs, the document number is returned.	Koopman (2014)
17	IRS	Retrieval judgement	IRS judges whether a document is be retrieved or not-retrieved	No change – IRS judges whether a document is be retrieved or not-retrieved	Langville and Meyer (2007)
18	Evaluation	Performance measurements	Term frequency ( <i>tf</i> ) is used to calculate the number of times a term occurs in a document	Phrase-term frequency ( <i>ptf</i> ) is used to calculate the number of times a phrase-term occurs in a document	Kang et al. (2015)
19	Evaluation	Performance measurements	Document frequency ( <i>df</i> ) is used to calculate the number of documents a term occurs in	Document frequency ( <i>df</i> ) is used to calculate the number of documents a phrase-term occurs in. This value is derived from <i>ptf</i> where $ptf > 0$ for a document.	Agnihotri et al. (2017)
20	Evaluation	Performance measurements	Collection frequency ( <i>cf</i> ) is used to calculate the number of times a term occurs in a collection.	Collection frequency ( <i>cf</i> ) is used to calculate the number of times a phrase-term occurs in a collection. This value is derived from <i>ptf</i> where $ptf > 0$ for a document.	Van Gysel et al. (2017)
21	Evaluation	Performance measurements	Precision (P), Recall (R) and F-measure (F) are used to calculate the effectiveness of the system.	Precision (P), Recall (R) and Specificity (S) are used to calculate the effectiveness of the system.	Cleverdon and Keen (1966); Choudhary et al. (2017)
22	Evaluation	Performance measurements	Inverse document frequency ( <i>idf</i> ) and <i>tf_idf</i> are used to weight the results from the impreciseness of the inverted indexing method	Inverse document frequency ( <i>idf</i> ) and <i>tf_idf</i> are unnecessary as weighting is not required owing to the exact matching capability of the hybrid indexing method	Spärck Jones (1972); Agnihotri et al. (2017)

### 5.3.1 The User stage

Referring back to Chapter Two where the original conceptual framework was presented, the user stage remains unchanged. The user stage begins when a user (a researcher) has an information need that he/she desires to be satisfied. The user then gathers all his/her documents collected over the years pertaining to the research topic and then reads through each document individually in the language of his/her choice. While perusing each document within this collection at some point the user makes a judgement as to whether the document is relevant (*tpfn*) or non-relevant (*fptn*) to his/her information need.

### 5.3.2 The IRS stage

By using DSR and pilot testing, the IRS a number of significant changes were made to the IRS stage. There are two main processes during the IRS stage: the information

gathering process and the search engine process (Manning et al., 2008). During the information gathering process, the IRS acquires the content from each document by: firstly using OCR software to convert the source file format to text file format, and secondly by transforming the text in the following ways:

- i) all (not a selected few) special characters (hyphens, punctuation, parenthesis) were replaced with a pipe delimiter (Harris, 2002) thus leaving only letters, numbers and the delimiter,
- ii) all the text was case folded to lowercase (Ruthven & Lalmas, 2003), and
- iii) chunks of text between the delimiters were acquired and stored as tokens together with their document identifiers and their unique Token ID (to maintain word proximity and ordinality) within the hybrid token index (Harris, 2002).

During the search engine process, phrase-terms are received from the user and these phrase-terms are then expressed as queries (expanded or not) as a representation of the user's information need. Once the query is presented to the search engine, the interrogation begins where attempts are made to match each phrase-term from the query stored in the hybrid query index, to the tokens existing within the hybrid token index. This process maintains word ordinality and proximity. The IRS then makes a judgement, whether to retrieve (*tpfp*) or not retrieve (*fnfn*) the document from the collection. During this judgment process, no ranking and weighting methods were utilised to influence the IRSs judgement. Only true values of *ptf* were used.

### 5.3.3 The Evaluation stage

The evaluation stage works with the judgement results made by the user (*tpfn* and *fpfn*) and the results made by the IRS (*tpfp* and *fnfn*). These results are dropped into a 2x2 contingency table (Luhn, 1953) and from this table, the values of *tp*, *fp*, *fn* and *tn* are derived. To measure the performance of the IRS, various formulae are available. The most commonly used formulae according to the literature in IRS evaluation are Precision and Recall (Cleverdon, 1956) and F-measure (Van Rijsbergen, 1979), but Specificity, although uncommon in the IRS literature (Choudhary et al., 2017), was required to test hypothesis **H3**. To compare two systems statistically the one-tailed student's t-test (Smucker et al., 2007) and the Kappa coefficient (Cohen, 1960) were used.

## 5.4 Findings: IRS-H versus IRS-I evaluation

During the evaluation between IRS-H and IRS-I, three hypotheses where tested, the findings of which are presented.

#### **5.4.1 Increasing effectiveness in retrieving relevant documents**

For hypothesis **H1**, the control group was IRS-I, and the test group IRS-H. The independent variable was 'hybridised indexing' and the dependant variable 'retrieval effectiveness'. The objective of the first hypothesis was to test whether an IRS using a hybrid indexing method increases the effectiveness of retrieving only those documents that are judged relevant by the user. Two IRSs, the first IRS-I using the inverted indexing method and the second IRS-H using the hybrid indexing method were evaluated for precision, average precision, MAP (Hardik & Jyoti, 2012), ranked and statistically analysed using a one-tailed student's t-test. From the results in Chapter Four, the alternative hypothesis **H1<sub>1</sub>** was accepted as the results were evident that  $MAP_{IRS-H}$  was greater than  $MAP_{IRS-I}$  with a statistical significance of  $p = 0.0365$  and therefore the first alternative hypothesis **H1<sub>1</sub>** held true (section 4.10).

***Finding 3:** Hybridised indexing increases the effectiveness of retrieving relevant documents*

#### **5.4.2 Reducing incorrect identification of relevant documents**

For hypothesis **H2**, the control group was IRS-I, and the test group IRS-H. The independent variable was 'hybridised indexing' and the dependant variable 'incorrect identification of relevant documents'. The objective of the second hypothesis was to test whether the hybrid indexing method reduces errors in incorrect identification of user judged relevant documents, thus reducing the number of documents for the user to peruse. IRS-I and IRS-H were evaluated for Recall (Langville & Meyer, 2007), average recall, MAR, ranked and statistically analysed using a one-tailed student's t-test. From the results in Chapter Four, the alternative hypothesis **H2<sub>1</sub>** was accepted as the results were evident that  $MAR_{IRS-H}$  was less than  $MAR_{IRS-I}$  with a statistical significance of  $p < 0.001$  and therefore the second alternative hypothesis **H2<sub>1</sub>** held true.

***Finding 4:** Hybridised indexing reduces the incorrect identification of relevant documents*

#### **5.4.3 Increasing quality in rejecting non-relevant documents**

For hypothesis **H3**, the control group was IRS-I, and the test group IRS-H. The independent variable was 'hybridised indexing' and the dependant variable the 'quality in rejecting non-relevant documents'. The objective of the third hypothesis was to test whether the hybrid indexing method increases the rejection quality of non-user relevant documents, thus providing confidence to the user in the judgement of the IRS. IRS-I and IRS-H were evaluated for specificity, average specificity, MAS

(Choudhary et al., 2017), ranked and statistically analysed using a one-tailed student's t-test. From the results in Chapter Four, the alternative hypothesis **H3<sub>1</sub>** was accepted as the results were evident that  $MAS_{IRS-H}$  was greater than  $MAS_{IRS-I}$  with a statistical significance of  $p < 0.001$  and therefore the third alternative hypothesis **H3<sub>1</sub>** held true.

**Finding 5:** *Hybridised indexing increases the quality in rejecting non-relevant documents*

### **5.5 Findings: IRS-H versus User judgements**

For hypothesis **H4**, the control group was the user and the test group IRS-H (Note that by definition user and IRS-H are both defined as systems). The independent variable was the 'hybridised indexing method' and the dependant variable 'agreement in judgements'. The objective of the fourth hypothesis was to determine whether the judgments made by the hybrid indexing method and the user agree. Disagreements between judgements of the users were found (Table 4.9). Strength of agreements ranged from slight to fair, to moderate, to substantial (Landis & Koch, 1977; Fleiss et al., 2003). All users missed matching phrase-terms to documents thus creating these disagreements. All users incorrectly stated at least five times, that a match did not occur between a phrase-term and a document, when it did. All users incorrectly stated a match occurred at least twice between a phrase-term and a document, when it did not match. Therefore, the fourth null hypothesis **H4<sub>0</sub>** held true.

**Finding 6:** *Judgments made by the hybrid indexing method and the user disagree*

**Finding 7:** *Judgments made between users disagree*

### **5.6 Findings: IRS-H versus User results**

For hypothesis **H5**, the control group was the user and the test group IRS-H. The independent variable was the 'hybridised indexing method' and the dependant variable 'satisfying the information needs of the user'.

The objective of the fifth hypothesis was to determine whether the results from the IRS using the hybrid indexing method satisfied the information needs (Singhal, 2001) of the user. The hybridised indexing method succeeded in exact matching of a query to a document, and succeeded in maintaining word ordinality and proximity. However, the users' results indicate that the hybrid indexing method did not satisfy the information needs of the user (Table 4.12). Therefore, the fifth null hypothesis **H5<sub>0</sub>** held true.

**Finding 8:** *The results generated by the hybrid indexing method and the user differ*

**Finding 9:** *The hybrid indexing method does not satisfy the information needs of the user*

### **5.7 The second research question**

The second research question is presented as:

**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

The second aim of this research, to enable the second research question to be answered, was to perform an explanatory study, via experimentation, to determine whether or not the design of the hybrid indexing method solved the problem of mismatched vocabulary between a query and a document. Referring to Figure 5.1, the processes to be followed to be in the position to answer the second research question are summarised:

- i) by comparing the IRS-I results to those of IRS-H: i) Precision increased thus increasing the effectiveness of retrieving relevant documents; ii) Recall decreased reducing incorrect identification of relevant documents; and iii) Specificity increased thus increasing the quality in rejecting non-relevant documents.
- ii) by comparing the results of IRS-H with the user: i) judgments made by the hybrid indexing method and the user disagreed; and ii) IRS-H did not satisfy the information needs of the user.

The second research question is discussed in further detail followed by the findings specified when and where appropriate. Finding 9 refers to the fact that phrase-terms must be appropriately specified by the user to meet the demands of the information need (Van Rijsbergen, 1979; Nguyen et al., 2018). If these phrase-terms are inappropriately specified, these demands will not be met and the results will not satisfy the information needs of the user. Although IRS-H performed exact matches of phrase-terms in documents, as specified by the user, IRS-H did not manage to satisfy the information needs of the user.

A query may be expanded to include multiple phrase-terms (Jimmy et al., 2018). These multiple phrase-terms are used to describe a topic in some way. In Pilot 3 (Appendix C), 14 phrase-terms were used to describe vocabulary mismatch in its many forms in an attempt to satisfy a set of 14 information needs, four of which were:



i) term mismatch; ii) vocabulary problem; iii) vocabulary gap; and iv) vocabulary mismatch (Liu et al., 2017).

If a search is triggered to find documents that relate to the topic 'vocabulary mismatch', all the phrases that describe the topic must be specified so that all documents relating to 'vocabulary mismatch' can be found. It is up to the user to specify the correct phrase-terms correctly within the query. At this point, the user becomes reliant on the IRS-H to be specific, to find the words expressed within the queries and to return the document. But because 'vocabulary mismatch' can be expressed in many ways ('vocabulary problem', 'vocabulary gap') all known phrases must exist in the query to retrieve the document sought (Liu et al., 2017). Hence, the user must already be knowledgeable about the topic to be able to do this (Soldaini et al. 2016; Shekarpour et al., 2017). It is up to the user to determine what the correct phrase-terms are and to express them within a query. The user must therefore do his/her homework making sure that all the correct phrase-terms relating to the topic are expressed (Van Gysel, 2017; Onal et al., 2018). Once the set of phrase-terms is complete, if the user uses the set and then searches for the topic using IRS-H, the documents will be returned based on an exact query/document match.

The results show that even though different phrase-terms were used and that these phrase-terms existed within the text of the document, on a few occasions the user still judged the document as non-relevant. This can occur when either an author uses a phrase-term within a document but the phrase-term did not relate to the topic of the document, or the phrase-term existed in the document but the user judged the document not relevant. Using an IRS that performs exact matching, some documents might still not be relevant to the user, because of how a user interprets them (Shekarpour et al., 2017).

Saving time by reading fewer documents is now possible. Referring to Appendix I, Table I.1 for IRS-I, and Table I.2 for IRS-H, the performance measurements for query  $q_{01}$  differed between systems. In this example, there were 100 documents in the collection. IRS-I read 100 documents ( $tpfpfn$ ) and retrieved 72 ( $tpfp$ ) documents. The user perused the documents, and judged 25 ( $tpfn$ ) of the 72 as relevant. For IRS-H 100 documents were read and 25 were retrieved. The user perused the document and judged 22 of the 25 as relevant. In this example, using IRS-I, the user was required to read 72 documents to find 25 as relevant. Using IRS-H, the user was required to read 25 documents to find 22 as relevant. By using IRS-H, time is thus saved for the user preventing the unnecessarily reading of approximately 200% more documents. This is owing to the increase in Specificity (Cleverdon & Keen, 1966; Dinh

& Tamine, 2015; Choudhary et al., 2017). The results show, in this example, that Specificity increased from 35% to 88% thus improving the quality in rejecting non-relevant documents.

***Finding 10:*** *The hybrid indexing method reduces research time by reading fewer non-relevant documents*

The user was supposed to be the 'known' from the results of the questionnaire so that IRS-H could be tested against the user. As in many IRS research projects, the goal is to try to get the IRS be as good as the user. However, the results show that IRS-H has become more effective than the user, since the user made mistakes in judgement (Landis & Koch, 1977; Goldstuck, 2019). Therefore, the user should not be used as the control group in all IRS experiments. The IRSs are becoming more and more effective and now challenge the role of the user.

***Finding 11:*** *IRSs are becoming more and more effective and now challenge the role of the user*

In past research of Cleverdon and Keen (1966), Manning et al. (2008), and Croft et al. (2015) it is suggested that the user must judge an IRS to determine its precision, recall, specificity, and its overall effectiveness. Past research suggests that questions should be posed, such as, what does the user think of the IRS? How does this IRS perform? Let a test be performed with the user to determine whether the IRS satisfies the information needs or nor. There are many unknowns here of how the user will interpret the data and make a judgement as it depends on the users perspective and what the user is looking for. If a user is asked to check that a phrase-term exists within a document and the user says 'yes' and the IRS says 'yes' then the results are acceptable as there is a 100% match. If the user says 'no' and the IRS 'yes' and the phrase-term exists in the document then the user is incorrect and the IRS correct. One may well ask: why blame the IRS if the resulting judgements of the IRS and user differ?

***Finding 12:*** *Why blame the IRS if the resulting judgements of the IRS and user differ?*

If the IRS and user disagree, which system is wrong and which system can you trust? If the user says 'no' and the IRS says 'no' and the phrase-term does not exist in the document, then the results are acceptable as there is a 100% agreement. At this point, the two systems (User and IRS-H) are just looking for agreement (Landis & Koch,

1977; Fleiss et al., 2003). If both systems agree that the phrase-term does not exist, then this is a good result and if both systems agree that the phrase-term does exist then this is also a good result. It is when these systems disagree that a problem is created. This potentially occurs when the design of the IRS disallows the matching of a phrase-term when it exists in a document, or when the user has missed identifying the existence of a phrase-term in a long document (Hiemstra, 2009; Narayan et al., 2017), or in a lengthy thesis, as evidenced in this research, and judges that the phrase-term does not exist when in fact it does.

***Finding 13:*** *Document length affects the judgement ability of the user*

IRS-H uses the design of an exact match between a query and a document (Pilot 3: Appendix C, section C.3, Table C.15). By using IRS-H index design, single or multi-term word search is now not a guess but an exact match (Van Rijsbergen, 1979; Van Gysel, 2017). That chance, or possibility, or probability is now eliminated. In all three pilots and during the evaluation exact matches occurred for all phrase-terms expressed in queries to those phrase-terms that existed within the documents.

***Finding 14:*** *During Pilot testing and evaluation the hybrid index design solved the vocabulary mismatch problem by matching queries to documents exactly*

To answer the second research question:

*The hybrid index design solves the vocabulary mismatch problem of matching a query to a document*

## **5.8 Summary of the framework**

The summary of the findings from the discussion revisiting the first research question, the hypotheses, IRS evaluations, IRS judgements, IRS results, and the second research question are collated and presented in Table 5.2.

Table 5.2: Summary of findings

Finding No.	Description
1	The hybrid indexing method maintains word ordinality and word proximity
2	The hybrid indexing method matches a phrase-term query to a document exactly
3	Hybridised indexing increases the effectiveness of retrieving relevant documents
4	Hybridised indexing reduces the incorrect identification of relevant documents
5	Hybridised indexing increases the quality in rejecting non-relevant documents
6	The judgments made by the hybrid indexing method and the user disagree
7	The judgments made between users disagree
8	The results generated by the hybrid indexing method and the user differ
9	The hybrid indexing method does not satisfy the information needs of the user
10	The hybrid indexing method reduces research time by reading fewer non-relevant documents
11	IRSs are becoming more and more effective and now challenge the role of the user
12	Why blame the IRS if the resulting judgements of the IRS and user differ?
13	Document length affects the judgement ability of the user
14	During Pilot testing and evaluation the hybrid index design solved the vocabulary mismatch problem by matching queries to documents exactly

To summarise the conceptual framework, in the hybrid indexing method, the token index is replaced by a pair of indices: i) the hybrid token index, and ii) the hybrid query index. During the search engine process the 'terms' expressed in the queries are replaced with 'phrase-terms' that enable the functionality of these new indices to be exercised. No weighting or ranking is necessary because of the exact matching technique used. In addition specificity, now well non-relevant documents are rejected, was an important aspect in this research, thus this measure is included in the framework. This conceptual framework can now be further tested and expanded by other researchers.

## CHAPTER SIX: CONCLUSIONS AND RECOMMENDATIONS

*“Great things are not done by impulse, but by a series of small things brought together.”*

– Vincent van Gogh

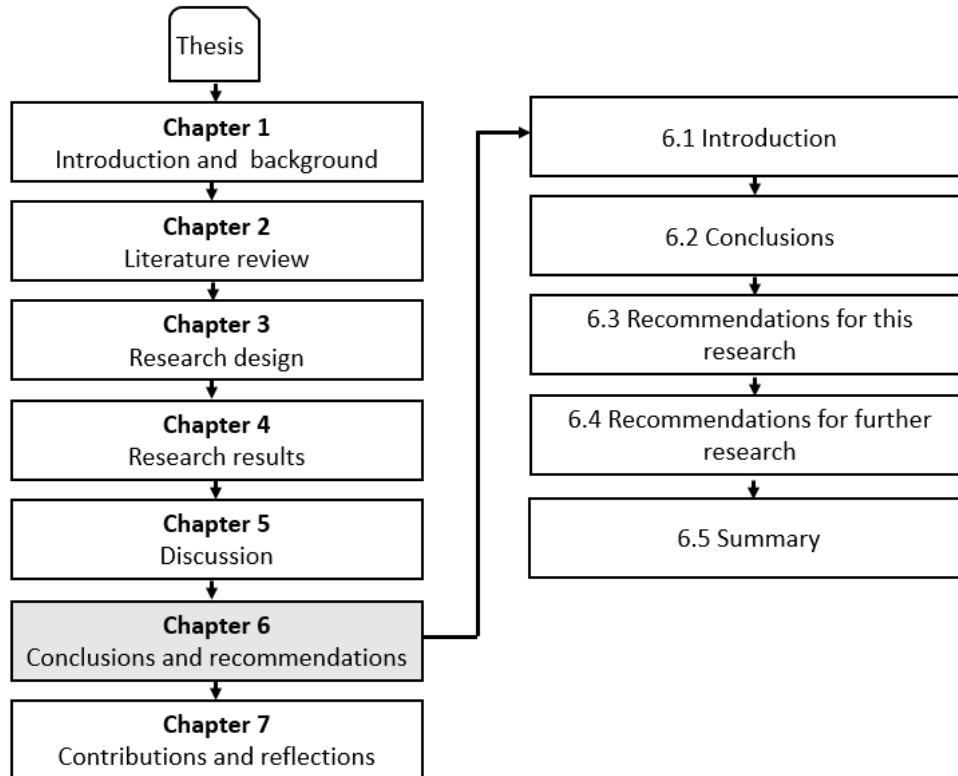


Figure 6.1: Schematic representation of Chapter Six

### 6.1 Introduction

Chapter Six presents the conclusion of this research together with recommendations.

### 6.2 Conclusions

The research questions, hypotheses and the aims of the study are now concluded.

#### 6.2.1 The first research question

The first research question posited in this study was:

**RQ1: How can an IRS index be designed that maintains word ordinality and word proximity?**

The first research question was answered by presenting a proposed framework of the hybrid indexing method and, based upon this method, by proving the design of IRS-H that utilised a pair of hybrid indices together with a unique token identity number that maintained word ordinality and proximity. Once rigorously pilot tested three times

and finally built, the hybrid indexing method was evaluated and shown to work through the use of statistical significance tests.

### **6.2.2 Hypotheses tested: IRS-H versus IRS-I**

Three hypotheses **H1**, **H2** and **H3** were posited in this study comparing the effectiveness of two systems: IRS-H versus IRS-I.

**H1<sub>0</sub>**: Hybridised indexing does not increase the effectiveness of retrieving relevant documents

The objective of the first hypothesis was to test whether an IRS using a hybrid indexing method increases the effectiveness of retrieving only those documents that are judged relevant by the user. By using mean average precision and by performing statistical significance tests using a one-tailed student's t-test the alternative hypothesis was accepted and therefore hybridised indexing increases the effectiveness of retrieving relevant documents.

**H2<sub>0</sub>**: Hybridised indexing does not reduce the incorrect identification of relevant documents

The objective of the second hypothesis was to test whether the hybrid indexing method reduces errors in incorrect identification of user judged relevant documents, thus reducing the number of documents for the user to peruse. By using mean average recall and by performing statistical significance tests using a one-tailed student's t-test the alternative hypothesis was accepted and therefore hybridised indexing reduces the incorrect identification of relevant documents.

**H3<sub>0</sub>**: Hybridised indexing does not increase the quality in rejecting non-relevant documents

The objective of the third hypothesis was to test whether the hybrid indexing method increases the rejection quality of user non-relevant documents, thus providing confidence to the user in the judgement of the IRS. By using mean average specificity and by performing statistical significance tests using a one-tailed student's t-test the alternative hypothesis was accepted and therefore hybridised indexing increases the quality in rejecting non-relevant documents.

### 6.2.3 Hypotheses tested: IRS-H versus User

Two hypotheses **H4** and **H5** were posited in this study that used a measure of the level of agreement between the judgements of the two systems: IRS-H versus the user.

**H4<sub>0</sub>**: Judgments made by the hybrid indexing method and the user disagree

The fourth hypothesis was tested by measuring the level of agreement between the user and IRS-H of whether a number of phrase-terms in search queries matched phrase-terms within documents. As there was a level of disagreement between the user and IRS-H, the judgments made by the hybrid indexing method and the user disagreed.

**H5<sub>0</sub>**: The hybrid indexing method does not satisfy the information needs of the user

The fifth hypothesis was tested by measuring the level of agreement between the user and by IRS-H of whether IRS-H satisfied the information needs of the user by retrieving those documents from the collection that were relevant to the user. As there was a level of disagreement between user and IRS-H, the hybrid indexing method did not satisfy the information needs of the user.

### 6.2.4 The second research question

The second research question posited in this study was:

**RQ2: Does the hybrid index design solve the vocabulary mismatch problem of matching a query to a document?**

In all three pilots and during the evaluation, exact matches occurred for all phrase-terms expressed in queries to those phrase-terms that existed within the documents. Therefore, the hybrid index design solves the vocabulary mismatch problem of matching a query to a document.

Revisiting the four statements of the research problem:

- i) *“Challenges exist for information retrieval systems in handling mismatching vocabularies in queries and candidate source documents”* (Onal et al., 2018).

These challenges are removed since vocabularies in queries can now match those in documents, when using the hybrid indexing method.

- ii) *“As a result these information retrieval systems may retrieve some documents that are non-relevant and miss some that are relevant” (Van Gysel, 2017).*

Using the hybrid indexing method allows the IRS to retrieve documents more effectively through increased precision and reduced recall. However, relevancy remains a judgement of the user.

- iii) *“This increases the time for research by forcing additional perusal of unsatisfactory results, and additional searches using alternative vocabularies” (Liu et al., 2017).*

Specificity increases the quality in rejecting non-relevant documents when using the hybrid indexing method. This increased quality provides the user with in the judgements made by the IRS and reduces the number of documents retrieved by the IRS that need to be perused, thus saving time.

- iv) *“This renders information retrieval systems less effective than they could be, and inhibits productive research” (Mitra & Awekar, 2017; Nguyen et al., 2018).*

Research becomes more productive as the hybrid indexing method is more effective in retrieving relevant documents. The user can be assured that when searching using a phrase-term the phrase-term will exist in all documents retrieved.

### **6.2.5 Aim of the study**

The first aim of this research was to perform an exploratory study, using DSR by designing, building and then rigorously pilot testing a new IRS using a hybrid indexing method that maintained word ordinality and word proximity, and to provide the conceptual framework for this method. This hybrid indexing method, IRS-H (Figure 6.2), was designed using a pair of indices that could maintain word ordinality and proximity successfully.



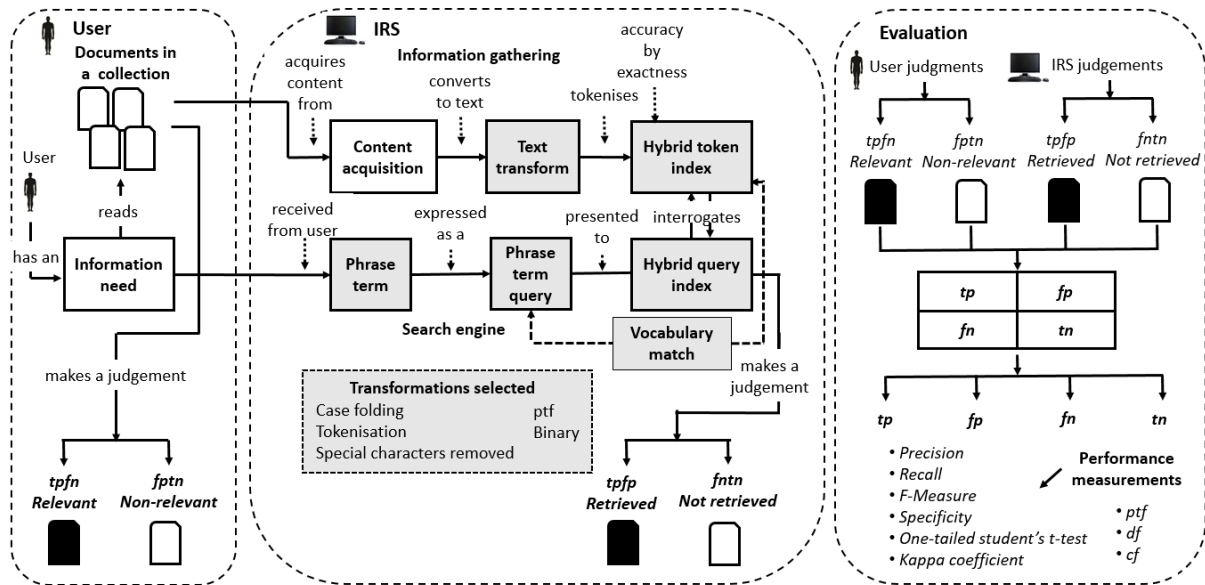


Figure 6.2: The proposed framework of the hybrid indexing method

The second aim was to perform an explanatory study, via experimentation, to determine whether the hybrid indexing method solved the problem of mismatched vocabulary between a query and a document. Based upon the two systems of IRS-H and IRS-I, the three objectives were:

- i) to test by hypothesis whether the effectiveness of retrieving those documents judged relevant by the user increased using IRS-H compared with using IRS-I. In this research IRS-H and IRS-I were successfully tested and comparisons made, the required data were successfully generated from both systems, and this data were statically analysed with significance;
- ii) to test by hypothesis whether the incorrect identification of relevant documents was reduced, by comparing IRS-H with IRS-I. In this research IRS-H and IRS-I were successfully tested and comparisons made, the required data were successfully generated from both systems, and this data were statically analysed with significance;
- iii) to test by hypothesis whether the quality in rejecting non-relevant documents was increased by comparing IRS-H with IRS-I. In this research IRS-H and IRS-I were successfully tested and comparisons made, the required data were successfully generated from both systems, and this data were statically analysed with significance.

These objectives were all achieved through the design, build, and pilot testing of IRS-H, and by statistical significance testing that proved IRS-H worked. This newly developed working method, that has numerous practical benefits in industry and

research (section 7.2.3), made significant progress in addressing the vocabulary mismatch problem and that of attempting to satisfy the information needs of the user. Based upon the two systems of IRS-H and the user, the objectives were:

- i) to test by hypothesis the judgements made between IRS-H and user to determine whether the judgements between IRS-H and the user agreed. In this research, the methods using IRS-H and user were successfully tested, the required data were successfully generated from both systems (IRS-H and user), and this data were statically analysed with significance; and
- ii) to test by hypothesis the judgements made between IRS-H and user to determine whether this hybrid indexing method satisfied the information needs of the user. In this research the methods using IRS-H and user were successfully tested, the required data were successfully generated from both systems (IRS-H and User), and this data were statically analysed with significance.

The final objective, to conclude the aim of this research, was:

- i) to determine whether IRS-H solved the vocabulary mismatch problem of matching a query to a document. This was concluded by drawing the results from the hybrid indexing method in the exploratory study and the results generated by the explanatory study.

In addition, the results from the statistical analysis in this research are not the contribution to knowledge as the statistics generated were used to prove the hybrid indexing method worked. In this research, the contribution to knowledge is the hybrid indexing method.

#### **6.2.6 General**

This research has evidenced that the hybrid indexing method can search for phrase-terms within the text of a document effectively. This method has the ability to search for one, two, three, four, or more words contained together within a phrase-term expressed as a query. By design, this method has the capability of extracting a whole paragraph from text which could be extremely useful when quoting references or searching for subject matter during academic research. In addition this method is not limited to the English language but can be used in any language that uses the 26 letter Latin based alphabet, and by those working in a variety of disciplines.

#### **6.2.7 Summary**

Information retrieval systems need to be more effective in order to assist those performing research in a better way. It is hoped that this contribution to knowledge,

the hybrid indexing method, has achieved this improvement, and will support academics, postgraduate researchers, attorneys, health carers, engineers, anthropologists, and many others.

### **6.3 Recommendations for this research**

The recommendations for this study are now concluded.

Referring to the design of the hybrid indexing method (section 5.3, Table 5.1) 22 findings were listed. These design findings compared IRS-H to IRS-I with ideas and concepts extracted from the literature, some of which were used and others not. The main recommendations are taken from these findings and presented as recommendations for this research

#### **6.3.1 From a user's perspective**

- i) Documents must be collected to provide a closed document collection.
- ii) Information needs must be provided by the user.
- iii) The user must judge whether a document is relevant or non-relevant.

#### **6.3.2 From an IRS perspective**

- iv) By design, the hybrid indexing method uses a pair of indices: a token index and a query index.
- v) Text must be case folded to lowercase with special characters removed.
- vi) Tokens must be non-distinct with unique token IDs – combined as a key makes them distinct.
- vii) The token index will expand containing more tokens than the inverted index, increasing storage space requirements.
- viii) Stemming, stopping, suffix stripping, vectors and other techniques are unnecessary owing to the exact matching capability of the hybrid indexing method.
- ix) A query populates the query index, which interrogates the token index and returns a result.
- x) Queries can contain single- or multi-word phrase-terms.
- xi) Words in phrase-terms are case folded to lowercase with special characters removed.
- xii) Phrase-terms are treated individually but the words within the phrase-terms are treated as a whole.
- xiii) A query attempts to match each phrase-term within the query index to a range of tokens within the token index.
- xiv) Word ordinality is maintained.

- xv) Word proximity is maintained.
- xvi) A query attempts to match all words in a phrase-term to a string of text in a document exactly. If an exact match occurs, the document number is returned.

### 6.3.3 From an evaluation perspective

- i) Phrase-term frequency (*ptf*) is used to calculate the number of times a phrase-term occurs in a document – this replaces term frequency (*tf*).
- ii) Document frequency (*df*) is used to calculate the number of documents a phrase-term occurs in. This value is derived from *ptf* where  $ptf > 0$  for a document.
- iii) Collection frequency (*cf*) is used to calculate the number of times a phrase-term occurs in a collection. This value is derived from *ptf* where  $ptf > 0$  for a document.
- iv) Precision (P), Recall (R) and Specificity (S) are used to calculate the effectiveness of the system.
- v) Inverse document frequency (*idf*) and *tf\_idf* are unnecessary, as weighting is not required owing to the exact matching capability of the hybrid indexing method.

## 6.4 Recommendations for further research

Referring to the key findings of this research (section 5.8, Table 5.2) the recommendations for further research are now concluded. In total, 14 findings were listed from this research and the main recommendations are taken from these findings and presented as recommendations for further research.

- i) The hybrid indexing method has the ability to match phrase-terms expressed within a query to those within a document exactly – This method should be used by those researchers and others that are in need of high precision, high specificity and highly effective searching using IRSs.
- ii) Search engines should have options that the user can use to ‘set’ certain working parameters to achieve a pure non-influenced search. For example, ignore parenthesis, ignore special characters, perform phrase-term exact matching, disallow synonyms, and remove weighting and ranking algorithms.
- iii) Judgments made between users disagree – The reasons why users make mistakes in judgements, and how these mistakes can be avoided by users, must be investigated and determined.
- iv) IRSs do not satisfy the information needs of the user – there is a need to better understand what it is that makes a user decide that a document does or does not meet his/her information need.

- v) IRSs are becoming more and more effective and now challenge the role of the user – The user needs to become better at determining what he/she really needs – the guess work of IRSs is now over, they can now make an exact match between query and document.
- vi) During Pilot testing and evaluation the hybrid index design solved the vocabulary mismatch problem of matching a query to a document. However, there remains a need to better understand how an IRS can be improved and how the user can become more expert and precise in choosing words for queries that are to be presented to the IRS.
- vii) In the literature, it is stated that the user is the definitive judge of the truth when measuring the effectiveness of an IRS. This research has evidenced that users' judgements are not always correct and that mistakes can be made in the cognitive process by the user when searching for a phrase within a document. Further research should explore how the user thinks and how they make decisions about a document's relevance and how a document satisfies information need.
- viii) Document length: during the experiment, each of the five users selected 20 documents of varying lengths from the closed collection. It became apparent during the experiment that reading 20 documents and answering the questionnaire in five hours was a tough task. The experiment therefore rolled over into the following working day and all questionnaires were duly completed. In lengthy documents, it was highly improbable that a user could have read every page to determine whether a phrase-term existed there. The largest document contained 100,625 words, which was a set of journal articles. In addition, there were two theses that contained 53,378 and 33,477 words. It was noted that the last user to select their set of 20 documents acquired the three largest documents. In his work, Cleverdon (1956) stated that an important control measure was to monitor the time it took for a user to index a document, as humans naturally judge things differently. In this case it was not the indexing that was time consuming, as these indexes were created programmatically, but it was the time it took for each user to identify the existence of phrase-terms within a document. This time varied greatly between users. A recommendation for further research, in this case, is that during these experiments, where users are required to search documents carefully, enough time must be allocated and the number of documents should be reduced. In hindsight using ten users (rather than five) and reading ten documents (rather than 20) each would have been more practical and more time efficient.

- ix) It would be interesting to determine the behaviour and the decision making process of the user, when making the judgment of whether a document is relevant or not. In the questionnaire, the user was asked to indicate which phrase-terms actually existed in each document and then, for the expanded queries, whether each document was relevant to the information need. The results show that there were 19 cases where the user judged the documents as relevant when none of the query's phrase-terms existed in the documents. A recommendation for further research is thus: perhaps this is where Artificial Intelligence (AI) (Jamilly, 2019) and knowledge based systems play a role? If the decision making process of the user can be better understood and this understanding is encapsulated within an AI system, then the knowledge gap in judgement between the IRS and the user could well be narrowed (Croft et al., 2015; Korda, 2019)?

## **6.5 Summary**

Chapter Six presented the conclusions of the study together with recommendations from the key findings for this research and further research.

## CHAPTER SEVEN: CONTRIBUTIONS AND REFLECTIONS

*"Things can change so fast on the internet" - Tim Berners-Lee*

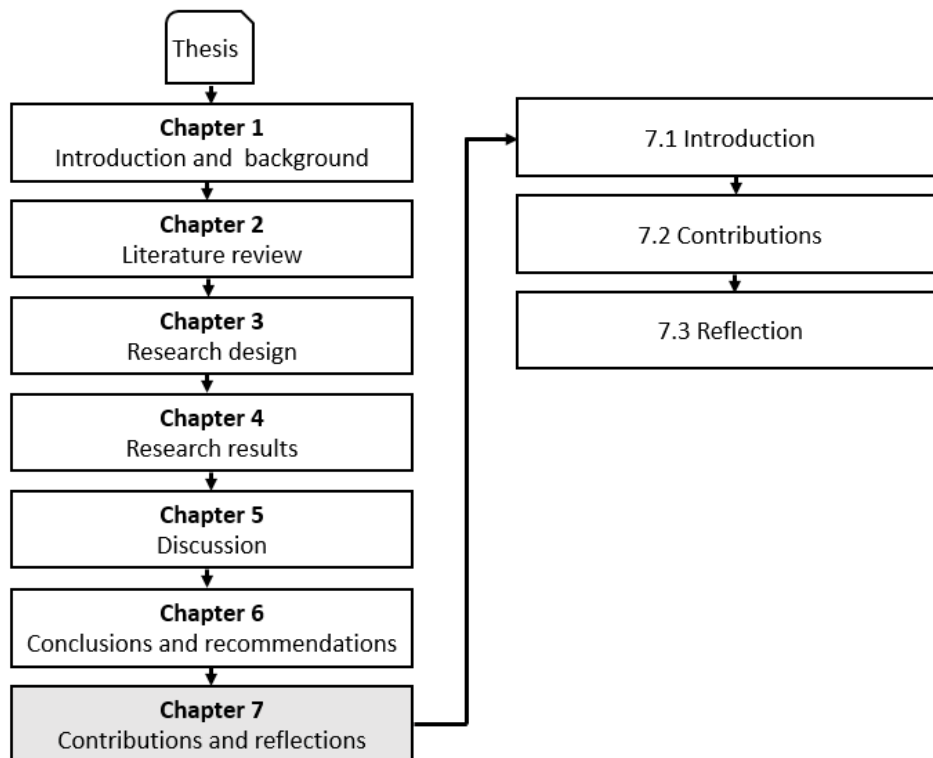


Figure 7.1: Schematic representation of Chapter Seven

### 7.1 Introduction

This chapter is presented in two sections: i) section 7.2 discusses the theoretical, methodological and practical contributions of this work; and ii) section 7.3 provides this researchers reflections on and assessment of the design, the contributions, the research method, the context of the research and finally some self-reflection.

### 7.2 Contributions

This section discusses the theoretical, methodological, and practical contributions of this work.

#### 7.2.1 Theoretical contributions

From this research, the contribution to knowledge is the hybrid indexing method, which is simultaneously a theoretical contribution and a methodological contribution (the theoretical design, which combines and extends many concepts from the literature). This method takes the inverted index and combines this method with the theoretical data retrieval property of exact matching; together with the key concept of the unique token identity number that maintains word ordinality and word proximity,

and uses the measurement of phrase-term frequency. The four theoretical contributions lay in the design concepts for the pair of indices: i) the hybrid token index; ii) the hybrid query index; iii) the unique token identity number; and iv) the phrase-term frequency. These design concepts are discussed in detail in Chapter Four and in Appendices A, B and C for the three pilots.

### **7.2.2 Methodological contributions**

Through the use of the hybrid indexing method the methodological contributions provide more effective retrieval of special-interest documents. This method accommodates mismatching vocabulary by using multiple synonymous phrase-terms, and it uses the concept of exact phrase matching to increase precision, to reduce recall, and to increase the quality of specificity. This method allows for expanded phrase-term queries (used to better describe a user's information need) and exact phrase matching (to better match a query to a document) to improve precision, reduce the retrieval of non-relevant documents, and increase the quality of rejected non-relevant documents. By design, this research provides a partial solution to a practical problem, by reducing the time required for the user to identify those documents relevant to his/her information need. This solution enables the user to perform multiple expanded phrase-term search queries and to retrieve more effectively the relevant documents within a shorter timeframe.

### **7.2.3 Practical contributions**

The hybrid indexing method can benefit many industries and different activities: postgraduate research, libraries, the motor industry, the legal profession, and information systems implementation are just some examples, but they are worth review in order to substantiate the claim of general benefits.

#### **7.2.3.1 Postgraduate research**

The problems of vocabulary mismatch became clear when this researcher attempted to retrieve appropriate documents that contained specific multi-word phrases during earlier postgraduate research. Writing a thesis can take many years and keeping track of the shifting literature is difficult. Trying to remember who said what, in which document (sometimes on which page), from a topic and referencing viewpoint was a challenge.

If you the reader are in research and you need a method to retrieve relevant documents with higher precision, and to retrieve fewer non-relevant documents thus saving precious perusal time, and need confidence in the quality, knowing that what was not retrieved is in fact truly non-relevant, then the hybrid indexing method should



be used. The title of this research is hybridised indexing for research based information retrieval and the word '*research*' in the title is used to exemplify the effectiveness and practicality of this hybrid indexing method to students and academics performing postgraduate research

### **7.2.3.2 The digital university library**

A university library can benefit from the use of the hybrid indexing method. Many university libraries currently use databases that link metadata to documents. These are often based on databases rather than IRSs and hence they use various fields that contain specific data for their search criteria. Using keywords to describe the content of a document is the best traditional method there is and has been in existence for many years, but has its limitations in search effectiveness and practicality. A major practical benefit to university students and academics would be to digitise a university's library and to convert and use text retrieval methodology rather than data retrieval methodology. If all the documents within a library were digitised, the hybrid indexing method could then be used to gather all the text from the library's documents and to store this text as tokens in the hybrid token index. Thereafter students and academics could use the search engine using the hybrid query index to retrieve relevant documents using expanded queries containing phrase-terms that match those that exist anywhere in the documents held within the university library's closed document collection. This method can already be used in any language using the 26 letter Latin based alphabet by people working in many disciplines, and its language inclusion can be extended.

### **7.2.3.3 The motor industry**

In the motor industry, brand names of motor manufacturers' products are expanding rapidly as new models are announced to meet clients' needs. Recently, Daimler AG announced new branding for their vehicles product range that now include: Mercedes Benz, Mercedes AMG and Mercedes Maybach. The hybrid indexing method enables effective retrieval of these product names presented as phrases in a query to the search engine as: 'mercedes benz' OR 'mercedes amg' OR 'mercedes maybach'. Motor dealers may also benefit from the increased effectiveness when searching for motor vehicles within their own company's closed document collections for: spare part material codes; searching using phrases that exist within the material descriptions; searching text in long description fields typically used in large software applications such as S4HANA and Maximo. Company names of motor vehicle manufacturers have also transformed over time owing to acquisitions and mergers. Take 'Daimler-Benz AG' for example, this company name has transformed into 'DaimlerChrysler AG' and more recently 'Daimler AG'. To search for this company, all three company names

can be utilised using the hybrid indexing method to extract the documents relevant to this company. What is essential however is for the user to understand how and when these transformations took place when searching aged archival documents: 'Daimler-Benz AG' was used from 1926 to 1998, 'DaimlerChrysler AG' from 1998 to 2007 and 'Daimler AG' from 2007 onwards. Multi-make and model naming conventions within the motor industry is another practical consideration. In efforts to increase efficiency and to reduce exhaust emissions, recent trends in the motor industry have been to downscale the cubic capacity of motor vehicle internal combustion engines and then turbocharge or supercharge them. In addition, electric vehicles are being brought to market and as a result, naming conventions to describe various models are changing. As traditional model names were cubic capacity based, for example, 5.0l V8, motor manufactures are changing their model names based on other criteria; for example, Jaguar Land Rover have now re-invented how they describe their models with the emergence of EVs using horse power output ratings; for example, 'Land Rover Range Rover Velar HSE P300' for a two litre four cylinder petrol engine producing 300 horse power and 'Jaguar I-Pace EV400 AWD HSE First Edition' for a pure electric vehicle with its electric motors producing 400 horse power. This makes searching for the correct vehicle more complex as there are many words/tokens in these phrases. The vocabulary mismatch problem arises when the same vehicle with the same specification receives a model name change, so for a user, wanting to effectively search will need to search for both the former and the current phrases that describe the motor vehicles correct naming convention for that specific make and model.

#### **7.2.3.4 The legal profession**

Within the legal profession, many large libraries of legal documents exist and if digitised can be accommodated by the hybrid indexing method. Searching for specific South African legal terms or the various Acts can be effectively searched, for example the Acts of: the *Money Bills Amendment Procedure and Related Matters Amendment Act 13 of 2018*, the *International Arbitration Act 15 of 2017* and the *Electoral Laws Amendment Act 1 of 2019*. These Acts can be expressed as queries utilising the phrases of: 'money bills amendment procedure and related matters amendment act 13 of 2018'; OR 'international arbitration act 15 of 2017' OR 'electoral laws amendment act 1 of 2019'. With the benefit of maintained word ordinality and word proximity of the hybrid indexing search engine, these phrases can be matched to the documents they reside in exactly. Thus, fewer documents are retrieved but these are the ones that are relevant to those working in the legal profession.

### 7.2.3.5 Information systems implementation

During the data migration process that is often necessary for new information systems implementation, metadata are often used to describe and index documents. Where large volumes of documents need to be migrated from one server to another server, metadata must be extracted to describe these documents. Often the first requirement is to determine what type of document it is (perhaps an invoice or a purchase order) and the second requirement is to find structure in the unstructured data (a customer number, an address, a company registration number). Refer to the Freeport-McMoRan/CMOC 'business world closed document collection' example in section 1.1.

## 7.3 Reflection

This section provides the researcher's reflections on and assessment of the design, contributions, the research method, the context of the research and finally some self-reflection.

### 7.3.1 Assessment of design

Reflecting back on this research, selecting DSR was an appropriate research methodology to design, build, and evaluate a purposefully built IRS. Before pilot testing began, many iterations of design and build took place using the deductive approach. The deductive approach was used to design the system and then after the results had been produced the deductive approach was used to modify and tweak any design flaws or programming inconsistencies. Many iterations of the design and build phases took place and once the system was stable pilot testing began.

The results from pilot testing brought a few surprises. Using Hamlet Act 3 Scene 1 written by William Shakespeare was ideal to test the systems using Elizabethan English / Early modern English where punctuation was abundant. This Pilot 1 helped in the early stages of the information gathering process and especially in the removal of special characters and the testing thereof. Pilot 2 used the book '*Ulysses*', written by James Joyce, this is where a few surprises arose. One assumption made during pilot testing was that a word could not be greater than 40 characters.

This assumption was proved incorrect when a 57-character word was acquired from the text:

*'handsomemarriedwomanrubbedagainstwidebehindinclonskeatram'.*

This necessitated a change in length of the token field within the hybrid token index. Using the subject of vocabulary mismatch for Pilot 3 helped in the testing of matching phrase-terms stored in the hybrid query index and matching those words with the

phrase-terms to the tokens stored in the hybrid token index. This test revealed a few important concepts pertaining to vocabulary mismatch that helped this researcher with the literature review of this thesis.

The IRS was built using a purposefully built Microsoft Access database (MS Access) and the Visual Basic (VB) programming language was used for coding. The database was adequate for this research but on a few occasions, especially when testing the IRS's information gathering process and the creation of the hybrid token index, MS Access reached its two gigabyte size limit. This necessitated special care and storage of this large index. The programming language was flawless and integrated well with the Windows version 10 (Windows) operating system, to enable the document files residing in Windows folders to be accessed and read reliably.

Although there was much effort involved in the design, build and evaluation the IRSs, much more effort was involved in the programming to produce the calculations and statistics required in order to evaluate the systems. For example the traditional term-by-document matrix used during the inverted index method evaluation contained 49 columns (one per term) times 100 rows (one per document) producing a matrix with 4,900 cells. As the hybrid indexing method used phrase-terms a phrase-term-by-document matrix was used to perform the evaluation containing 65 columns (one per phrase-term) times 100 rows (one per document) producing a matrix with 6,500 cells. These matrices cells stored the term frequencies, the number of times a term occurred in a document and the phrase-term frequencies, the number of times a phrase-term occurred in a document. These *tf* and *ptf* values were then converted to binary and stored in a second matrix for further statistical use. All the calculations required for Precision, Recall, and Specificity, their rankings, averages, and mean values, were programmatically calculated using the VB programming language. To perform this type of research, the researcher must have good skills in database design, in programming and in the use of structured query language (SQL) to produce these extensive calculations and statistics required to perform the evaluations of the IRSs.

Robertson (1981) argued that when new systems were developed, and the inadequacies of the old systems were revealed, these inadequacies would be replaced by new challenges arising from the newer system. This statement by Robertson did apply to this research as inadequacies found in previous indexing methods, resolved in this research, did bring new challenges. For example: (i) the size of the indices increased to accommodate the additional data stored within them, and (ii) the search response time slowed when the query index interrogated the token index.

The pair of indices, the hybrid token index and the hybrid query index, and the use of these as a method, are the principle design outcomes of this research. These need to be further tested. Simultaneously, this research has revealed this hybrid indexing method increases Precision, reduces Recall, and increases Specificity. This method and these results can now progress into further positivistic work. Digitising a university library's documents, testing this hybrid indexing method using these documents, and selecting a larger group of participating users would be an ideal further contribution to this type of research.

### **7.3.2 Assessment of contributions**

This research provided a proposed framework for the hybrid indexing method. This method used the novel design of a pair of indices, which utilised the concept of the unique token identity number, in the design of the hybrid token index and the hybrid query index, and the concept of phrase-term frequency as a measurement. This hybrid indexing method was evaluated and proved to work with statistical significance, and when compared with the traditional inverted index, performed better with increased effectiveness. Improvements in IRS effectiveness were achieved through the use of the hybrid indexing method, and the vocabulary mismatch problem between a query and a document was solved. However, the information needs of the user were not satisfied (this appears to be a human problem rather than an IRS problem).

### **7.3.3 Assessment of the research**

During the early stages of this study, the research strategy was mulled over, moving from case study research, ethnographic research, grounded theory, action research, action design research, case study research, and DSR. Choosing the appropriate research strategy for this research took a while, as it was different from traditional research in many ways. Although a problem existed and a solution was sought, an artefact had to be designed and built to help try to solve the problem. It was not until a workshop at CPUT between postgraduate university staff and Masters and Doctoral research students that the research strategy of DSR become well understood. Only after initial investigations through discussions, reading the specific literature intensely and playing video footage of DSR methods on the Web, did the pieces of the puzzle begin to fit together, and thereafter DSR was chosen as the appropriate strategy for this research. From the literature, the two most influential authors in design science that had a substantial impact on this research were Shirley Gregor and Alan Hevner (Gregor & Hevner, 2016) and in mathematics were Amy Langville and Carl Meyer (Langville & Meyer, 2007).

#### **7.3.4 Assessment of the context and research purpose**

The purpose of this study was to develop artefacts to solve the research problem by combining the results of two studies, one an exploratory study, where the hybridised indexing method was designed, built, and rigorously tested, and the second, an explanatory study based on hypotheses and the quantitative method to prove that this method worked. The built IRS-H using the hybrid indexing method worked, thus achieving the aims, objectives and the purpose of this research.

#### **7.3.5 Self-reflection**

This section presents this researcher's personal reflections on, and experience in, performing this study.

Fourteen years ago, a young member of this researcher's family fell ill with an incurable deadly disease of unknown cause. As years passed the family, with information needs to better understand the disease, began searching for answers making use of various IRSs including search engines. The crux of the problem was how to describe the disease correctly when searching for the literature as the disease formed part of a group of diseases with many classifications and synonymic nomenclatures. In addition, the less prevalent disease, and the one sought, affecting 10% of those diagnosed was difficult to separate, using queries and IRSs, from the more prevalent one resulting in large numbers of non-relevant and unwanted documents from the literature. During this period, the quality of relevant documents retrieved by the IRSs was poor. What was retrieved and judged as relevant by the IRS was not judged relevant by this researcher. Most documents retrieved related to the more prevalent disease. Because the disease had many classifications finding documents that matched the words expressed in the queries exactly for just the one classification was a challenge. This challenge laid the foundation for this experimental study, and hence the title: hybridised indexing for research based information retrieval. The most inspirational author in the early days of this research was Calvin Northrup Mooers (1950; 1951) as in his work he introduced the phrase '*descriptor*' the origin of the query term and it is he who coined the phrase '*information retrieval*'.

This research was driven by the need to find health care phrases effectively and prompted the design of the hybrid indexing method which was pilot tested over a few years. The hybrid indexing method works and has the ability to match words in phrases expressed by queries to those words in documents exactly, maintaining word ordinality and word proximity.

The learning experience from this research, triggered by the earlier personal event, was comprehensive. The problem of vocabulary mismatch was personally experienced which brought huge frustration in identifying documents that were relevant to the specific disease. Having this frustration, and because a system was not freely available to perform effectively what was required, a prototype IRS was developed, in an attempt to solve the problem of mismatching vocabulary. After a few iterative design cycles and through thorough testing the hybrid indexing method functioned well and produced the results hoped for from its design. The results were impressive, as those documents required to satisfy an information need, were system retrieved as relevant, from the author's personal collection of documents. In addition, and more importantly, the unwanted documents, pertinent to the other similar diseases, were system rejected with high quality, thus saving time in not reading the irrelevant documents.

Coming from an engineering and information technology background, the concept of information retrieval, transformation and indexing was well known. What was interesting was how the judgements by the users' were made, how the judgements by IRS-H were made, and how these differed and why. Much was learned from the comprehensive literature review and especially how the historical methods that were used before the advent of computers. The literature was not very explicit in how an IRS actually worked and therefore one of the outcomes of this research was to be able to explain just that. The many formulae and matrices used to produce the mathematical and statistical results were challenging. Therefore, programs were written to programmatically replicate the calculations using these formulae, and database tables were used to replicate the matrices and their functionality. This was necessary to enable repeated replication of these processes during the design cycles. It is hoped, that what is presented in this thesis will: (i) lay a foundation to those interested in information retrieval research, (ii) help those researchers, parents and family members to search for, and find specifically what they are looking for, precisely and effectively, and (iii) provide the necessary formulae, matrices and methods, required to design and evaluate an IRS.

**REFERENCE LIST**

*This list of references applies to references in both Volume I and Volume II of this thesis.*

ACM see Association for Computing Machinery.

Adouane, W. & Dobnik, S. 2017. Identification of languages in Algerian Arabic multilingual documents. *Proceedings. The 3<sup>rd</sup> Arabic Natural Language Processing Workshop (WANLP)*, Valencia, Spain, 3 April: 1-8.

Agnihotri, D., Verma, K. & Tripathi, P. 2017. An empirical study of clustering algorithms to extract knowledge from PubMed articles. *Transactions on Machine Learning and Artificial Intelligence*, 5(3):13-27.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716-723.

Ali, M.M. 2013. Mixed-language Arabic-English information retrieval. Unpublished PhD thesis, University of Cape Town, Cape Town.

Allones, J.L., Martinez, D. & Taboada, M. 2014. Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. *Journal of Medical Systems*, 38(134):1-14.

Alonso, O., Gertz, M. & Baeza-Yates, R. 2009. Clustering and exploring search results using timeline constructions. *Proceedings. The 18<sup>th</sup> ACM Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, China, 2-6 November: 97-106.

Andersson, L., Rekabsaz, N. & Hanbury, A. 2017. *Automatic query expansion for patent passage retrieval using paradigmatic and syntagmatic information*. Information and Software Engineering Group. <https://www.semanticscholar.org/paper/Automatic-Query-Expansion-for-Patent-Passage-using-Andersson-Rekabsaz/39e43cc740c44caa4bfc2aa4efcfdcabb8f882cd> [Accessed: 25 January 2018].

Anon. 2006. Introduction to Dialog for information professionals. *Tomson*, [https://tefkos.comminfo.rutgers.edu/courses/e530/readings/dialog\\_2006\\_intro\\_for\\_infopros.pdf](https://tefkos.comminfo.rutgers.edu/courses/e530/readings/dialog_2006_intro_for_infopros.pdf) [Accessed: 1 November 2019].

Association for Computing Machinery. 2019a. *ACM digital library*. <https://dl.acm.org/> [Accessed: 14 August 2019].

Association for Computing Machinery. 2019b. *Table of contents for issues of Communications of the ACM*. <http://ftp.math.utah.edu/pub/tex/bib/toc/cacm2010.html> [Accessed: 14 August 2019].

Ayumba, E.M. 2015. Modelling software agents: Web-based decision support system for malaria diagnosis and therapy. *Journal of Health Informatics in Africa*, 3(1):30-36.

Babbie, E. 2013. *The practice of social research*. 13<sup>th</sup> ed. Belmont, CA: Wadsworth.

Baskerville, R.L., Kaul, M. & Storey, V.C. 2017. Establishing reliability in design science research. *Proceedings. The 38<sup>th</sup> International Conference on Information Systems (ICIS 2017)*, Seoul, South Korea, 10-13 December.



- Baxter, I., Ouzzani, M., Orcun, S., Kennedy, B., Jandhyala, S.S. & Salt, D.E. 2007. Purdue Ionomics information management system. An integrated functional genomics platform. *Plant Physiology*, 143:600-611.
- Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53(1763):370-418.
- Bell, T.C., Moffat, A., Nevill-Manning, C.G. & Witten, I.H. 1993. Data compression in full-text retrieval systems. *Journal of the American Society for Information Science*, 44(9):508-531.
- Berners-Lee, T. 1989. Information management: a proposal. Unpublished proposal, European Organisation for Nuclear Research, Geneva.
- Berners-Lee, T. 2000. *Weaving the Web: origins and future of the World Wide Web*. London: Texere Publishing.
- Berners-Lee, T. & Fischetti, M. 1999. *Weaving the Web*. <http://www.w3.org/people/berners-lee/weaving/overview.html> [Accessed: 20 August 2019].
- Berners-Lee, T. & Fischetti, M. 2000. *Weaving the Web: the original design and ultimate destiny of the World Wide Web*. New York: Harper Business.
- Berry, M.W., Drmac, Z. & Jessup, E.R. 1999. Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2):335-362.
- Bi, K., Ai, Q. & Croft, W.B. 2019. Iterative relevance feedback for answer passage retrieval with passage-level semantic match. *Proceedings*. The 2019 European Conference on Information Retrieval (ECIR 2019), Cologne, Germany, 14-18 April: 558-572.
- Binkley, D. & Lawrie, D. 2015. The impact of vocabulary normalisation. *Journal of Software: Evolution and Process*, 27(4):255-273.
- Blair, D.C. & Maron, M.E. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289-299.
- Bonello, M. & Meehan, B. 2019. Transparency and coherence in a doctoral study case analysis: reflecting on the use of NVivo within a 'framework' approach. *The Qualitative Report*, 24(3):483-498.
- Brin, S. & Page, P. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1-7):107-117.
- Brudfors, M., Balbastre, Y. & Ashburner, J. 2019. Nonlinear Markov random fields learned via backpropagation. Unpublished paper, Cornell University, Ithaca, NY.
- Bui, D.D.A., Jonnalagadda, S. & Del Fiore, G. 2015. Automatically finding relevant citations for clinical guideline development. *Journal of Biomedical Informatics*, 57:436-455.
- Burrell, G. & Morgan, G. 1979. *Sociological paradigms and organisational analysis*. London: Heinemann.
- Bush, V. 1945. As we may think. *The Atlantic Monthly*, 3(2):35-46.
- Case, D.O. 2002. *Looking for information - A survey of research on information seeking, needs, and behavior*. Lexington, KY: Academic Press.

- Cape Peninsula University of Technology. 2019. *Ethics documentation for the Faculty of Informatics and Design*. <http://www.cput.ac.za/research-technology-and-innovation/ethics> [Accessed: 24 April 2019].
- Castells, P., Fernandez, M. & Vallet, D. 2007. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):261-272.
- Chang, C., Kayed, M., Girgis, M.R. & Shaalan, K. 2006. A survey of Web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411-1428.
- Chaparro, O., Florez, J.M. & Marcus, A. 2016. On the vocabulary agreement in software issue descriptions. *Proceedings*. The 2016 IEEE International Conference on Software Maintenance and Evolution (ICSME 2016), Raleigh, NC, USA, 2-7 October.
- Chen, Y. & Welling, M. 2012. Bayesian structure learning for Markov random fields with a spike and slab prior. *Proceedings*. The 28<sup>th</sup> Conference on Uncertainty in Artificial Intelligence (UAI 2012), Catalina Island, CA, USA, 14-18 August.
- Chew, P.A., Bader, B.W., Helmrich, S., Abdelali, A. & Verzi, S.J. 2011. An information-theoretic, vector-space-model approach to cross-language information retrieval. *Natural Language Engineering*, 17(1):37-70.
- Chitu, A. 2010. *Google's AROUND operator for proximity Search*. Google Operating System. <http://googlesystem.blogspot.co.za/2010/12/googles-around-operator.html> [Accessed: 20 August 2019].
- Choudhary, P., Kumar, S., Bachhawat, A.K. & Pandit, S.B. 2017. CSmetaPred: a consensus method for prediction of catalytic residues. *BMC Bioinformatics*, 18(583):1-13.
- Chunara, R., Freifeld, C.C. & Brownstein, J.S. 2012. New technologies for reporting real-time emergent infections. *Parasitology*, 139(14):1-14.
- Clarke, C.L.A., Cormack, G.V. & Tudhope, E.A. 2000. Relevance ranking for one to three term queries. *Information Processing and Management*, 36(2):291-311.
- CLEF see Cross-Language Education and Function.
- Cleverdon, C.W. 1956. Proposals for an investigation into the efficiency of various retrieval systems. Unpublished report, Cranfield University, Cranfield.
- Cleverdon, C.W. 1960. ASLIB Cranfield research project: report on the first stage of an investigation into the comparative efficiency of indexing systems. Unpublished report, Cranfield University, Cranfield.
- Cleverdon, C.W. 1967. The Cranfield tests on index language devices. Unpublished paper, Cranfield University, Cranfield.
- Cleverdon, C.W. 1991. The significance of the Cranfield tests on index languages. *Proceedings*. The 14<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1991), Chicago, IL, USA, 13-16 October: 3-12.

- Cleverdon, C.W. & Keen, M. 1966. ASLIB Cranfield research project – factors determining the performance of indexing systems, Volume 2, Test results. Unpublished report, Cranfield University, Cranfield.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37-46.
- Conger, A.J. 2017. Kappa and Rater accuracy: paradigms and parameters. *Educational and Psychological Measurement*, 77(6):1019-1047.
- CPUT see Cape Peninsula University of Technology.
- Creswell, J.W. 2013. *Research design: qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.
- Croft, W.B. 2019. The importance of interaction for information retrieval. *Proceedings*. The 42<sup>nd</sup> International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), Paris, France, 21-25 July.
- Croft, W.B., Metzler, D. & Strohman, T. 2015. *Search engines: information retrieval in practice*. Harlow: Pearson Education.
- Croft, W.B., Turtle, H.R. & Lewis, D.D. 1991. The use of phrases and structured queries in information retrieval. *Proceedings*. The 14<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1991), Chicago, IL, USA, 13-16 October: 32-45.
- Cross-language education and function. 2016. *Cross-language education and function*. CLEF. <http://www.clef-campaign.org> [Accessed: 15 August 2019].
- De Raadt, A., Warrens, M.J., Bosker, R.J. & Kiers, H.A.L. 2019. Kappa coefficients for missing data. *Educational and Psychological Measurement*, 79(3):558-576.
- DeLone, W.H. & McLean, E.R. 1992. Information systems success: the quest of the dependent variable. *Information Systems Research*, 3(1):60-95.
- Dervin, B. 1992. From the mind's eye of the user: the sense-making qualitative-quantitative methodology. Unpublished paper, Ohio State University, Columbus, OH.
- Dietz, L., Xiong, C., Dalton, J. & Meij, E. 2019. Special issue on knowledge graphs and semantics in text analysis and retrieval. *Information Retrieval Journal*, 22(3-4):229-231.
- Ding, C., Li, T. & Peng, W. 2006. Nonnegative matrix factorisation and probabilistic latent semantic indexing: equivalence, chi-square statistic, and a hybrid method. *Proceedings*. The 21<sup>st</sup> National Conference on Artificial Intelligence and the 18<sup>th</sup> Innovative Applications of Artificial Intelligence Conference (AAAI-06), Boston, MA, USA, 16-20 June.
- Dinh, D. & Tamine, L. 2015. Identification of concept domains and its application in biomedical information retrieval. *Information Systems and e-Business Management*, 13:647-672.
- Dougherty, C. 2019. *Statistical tables*. <https://home.ubalt.edu/ntsbarsh/business-stat/statisticaltables.pdf> [Accessed: 14 April 2019].

- EBSCO. 2019. How do I create a proximity search? *EBSCO*, [https://connect.ebsco.com/s/article/how-do-i-create-a-proximity-search?language=en\\_us](https://connect.ebsco.com/s/article/how-do-i-create-a-proximity-search?language=en_us) [Accessed: 1 November 2019].
- Egghe, L. 2008. The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Information Processing and Management*, 44(2):856-876.
- Egoli, O., Markovitch, S. & Gabrilovich, E. 2000. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems*, 1-38.
- Elragal, A. & Haddara, M. 2019. Evaluation in the lens of big data analytics. *Systems*, 7(2):1-8.
- Faheem, M. 2014. Intelligent content acquisition in Web archiving. Unpublished PhD thesis, Paris Institute of Technology, Paris.
- Faloutsos, C. & Jagadish, H.V. 1992. Hybrid index organisations for text databases. In Pirotte, A., Delobel, C. & Gottlob, G. (eds.), *Advances in Database Technology – EDBT '92*. EDBT 1992. Lecture Notes in Computer Science, Volume 580. Berlin: Springer.
- Farwick, M., Breu, R., Hauder, M., Roth, S. & Matthes, F. 2013. Enterprise architecture documentation: empirical analysis of information sources for automation. *Proceedings*. The 46<sup>th</sup> Hawaii International Conference on System Sciences, Hawaii, USA, 7-10 January: 3868-3877. IEEE.
- Fisher, R.A. 1971. *The design of experiments*. 8<sup>th</sup> ed. New York, NY: Hafner Publishing.
- Fleiss, J.L., Levin, B. & Paik, M.C. 2003. *Statistical methods for rates and proportions*. 3<sup>rd</sup> ed. Chichester: John Wiley & Sons.
- Frej, J., Chevallet, J.P. & Schwab, D. 2018. *Enhancing translation language models with word embedding for information retrieval*. arXiv:1801.03844 [cs.IR], Cornell University, Ithaca, NY.
- Furnas, G.W., Landauer, T.K., Gomez, L.M. & Dumais, S.T. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964-971.
- Gacenga, F., Cater-Steel, A., Toleman, M. & Tan, W.G. 2012. A proposal and evaluation of a design method in design science research. *Electronic Journal of Business Research Methods*, 10(2):89-100.
- Garfield, E. 1955. Citation indexes to science: a new dimension in documentation through association of ideas. *Science*, 122(3159):108-111.
- Garfield, E. 1972. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471-479.
- Garfield, E. 1997. A tribute to Calvin N. Mooers, a pioneer of information retrieval. *The Scientist*, 11(6):9-11.
- Garfield, E. 2007. The evolution of the Science Citation Index. *International Microbiology*, 10:65-69.

Goeuriot, L., Jones, G.J.F., Kelly, L., Muller, H. & Zobel, J. 2016. Medical information retrieval: introduction to the special issue. *Information Retrieval*, 19:1-5.

Goldstuck, A. 2019. Oracle to bring war on Amazon to SA shores. *Business Times*. 5. September 22. 00:07.

Gray, F. 1947. Pulse Code Communication. Unpublished patent, Bell Telephone Laboratories, New York.

Gregor, S. 2006. The nature of theory in information systems. *MIS Quarterly*, 30(3):611-642.

Gregor, S. 2014. *Theory construction*. Information systems symposium (ISS-2014). <https://www.youtube.com/watch?v=og08LGVjLWk> [Accessed: 7 November 2016].

Gregor, S. & Hevner, A.R. 2013a. Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 27(2):337-355.

Gregor, S. & Hevner, A.R. 2013b. Positioning and presenting design science research for maximum impact – Appendices. *MIS Quarterly*, 27(2):A1-A6.

Gregor, S. & Hevner, A. 2016. The digital innovation design activities wheel. *Presentation*. SIGPrag workshop on Practice-based Design and Innovation of Digital Artefacts, Dublin, Ireland, December.

Gregor, S. & Jones, D. 2007. The anatomy of a design theory. *Journal of the Association for Information Systems*, 8(5):313-355.

Gregor, S., Maedche, A., Morana, S. & Schacht, S. 2016. Designing knowledge interface systems: past, present, and future. In Parsons, J., Tuunanen, T., Venable, J.R., Helfert, M., Donnellan, B. & Kenneally, J. (eds.), *Breakthroughs and emerging insights from ongoing design science projects. Proceedings. The 11<sup>th</sup> International Conference on Design Science Research in Information Systems and Technology (DESRIST 2016), St. John, Canada, 23-25 May: 43-50*.

Gross, P.L.K. & Gross, E.M. 1927. College libraries and chemical education. *Science*, 66(1713):385-389.

Guba, E.G. & Lincoln, Y.S. 2005. Paradigmatic controversies, contradictions, and emerging confluences. In Denzin, N.K. & Lincoln, Y.S. (eds.), *Sage handbook of qualitative research*. 3<sup>rd</sup> ed. Thousand Oaks, CA: Sage: 163-188.

Gugnani, S. & Roul, R.K. 2014. Triple indexing: an efficient technique for fast phrase query evaluation. *International Journal of Computer Applications*, 87(13):9-13.

Gupta, C. 2008. Efficient k-word proximity search. Unpublished MSc thesis, Case Western Reserve University, Cleveland, OH.

Gusfield, D. 1997. *Algorithms on strings, trees and sequences: computer science and computational biology*. New York, NY: Cambridge University Press.

Ha, L.Q., Sicilia-Garcia, E.I., Ming, J. & Smith, F.J. 2002. Extension of Zipf's law to words and phrases. *Proceedings. The 19<sup>th</sup> International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, 26-30 August: 1-6*.

- Halácsy, P. & Trón, V. 2007. Benefits of resource-based stemming in Hungarian information retrieval. *Lecture Notes in Computer Science (LNCS)*, 4730:99-106.
- Hamid, A. 2017. Relevance feedback in information retrieval systems. Unpublished paper, Bahria University, Islamabad.
- Hamilton, L., Koehler, F. & Moitra, A. 2017. Information theoretic properties of Markov random fields, and their algorithmic applications. *Proceedings. The 31<sup>st</sup> Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 4-9 December.
- Hanbury, A., Lupu, M., Kando, N., Diallo, B. & Adams, S. 2014. Guest editorial: Special issue on information retrieval in the intellectual property domain. *Information Retrieval*, 17(5-6): 407-411.
- Hanid, M. 2014. Design science research as an approach to develop conceptual solutions for improving cost. Unpublished PhD thesis, University of Salford, Manchester.
- Huang, S. & Huang, Q. 2016. A hybrid index for characterising drought based on a nonparametric kernel estimator. *Journal of Applied Meteorology and Climatology*, 55:1377-1389.
- Hardik, J.J. & Jyoti, P. 2012. *Evaluation of some information retrieval models for Gujarati ad hoc monolingual tasks*. arXiv:1209.0126v1 [cs.IR]. Cornell University, Ithaca, NY.
- Harris, Z.S. 1954. Distributional structure. *Word*, 10(23):146-162.
- Harris, Z.S. 1976. A theory of language structure. *American Philosophical Quarterly*, 3(4):237-255.
- Harris, Z.S. 1979. Mathematical analysis of language. *Proceedings. The 6<sup>th</sup> International Congress of Logic, Methodology and Philosophy of Science*, Hannover, Germany, 22-29 August: 623-637. International Union of History and Philosophy of Science.
- Harris, Z.S. 2002. The structure of science information. *Journal of Biomedical Informatics*, 35(4):215-221.
- Hauff, C. 2010. Predicting the effectiveness of queries and retrieval systems. Unpublished PhD thesis, University of Twente, Enschede.
- He, B. & Ounis, I. 2009. Studying query expansion effectiveness. *Proceedings. The 31<sup>st</sup> European Conference on IR Research on Advances in Information Retrieval (ECIR)*, Toulouse, France, 6-9 April 2009.
- Henderson, M.M. 1967. *Evaluation of information systems: A selected bibliography with informative abstracts*. Washington, DC: US Government Printing Office.
- Heron, J. 1996. *Co-operative inquiry: research into the human condition*. London: Sage.
- Hevner, A.R. 2007. A three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2):87-92.
- Hevner, A.R. 2015a. *Is design science the future of innovation?* <https://www.youtube.com/watch?v=llsXxyiiQo> [Accessed: 4 November 2016].

- Hevner, A.R. 2015b. Designing informing systems: what research tells us. *Keynote*. Informing Science and IT Education Conferences (InSITE 2015), USA, Tampa, FL, USA, 29 June-5 July 2015.
- Hevner, A., Vom Brocke, J. & Maedche, A. 2019. Roles of digital innovation in design science research. *Business & Information Systems Engineering*, 61(1):3-8.
- Hevner, A.R., March, S.T., Park, J. & Ram, S. 2004. Design science in information systems research. *MIS Quarterly*, 28(1):75-105.
- Hiemstra, D. 2000. A probabilistic justification for using tf\*idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131-139.
- Hiemstra, D. 2009. Information retrieval models. In Goker, A. & Davies, J. (eds.), *Information retrieval: searching in the 21<sup>st</sup> Century*. Chichester: John Wiley & Sons.
- Hudson, D.A. & Manning, C.D. 2019. Learning by abstraction: the neural state machine. Unpublished paper, Cornell University, Ithaca, NY.
- IJzereef, L., Kamps, J. & De Rijke, M. 2005. Biomedical retrieval: how can a thesaurus help? In Meersman, R. & Tari, Z. (eds.), *On the move to meaningful internet systems 2005: CoopIS, DOA, and ODBASE*. OTM 2005. Lecture Notes in Computer Science, Volume 3761. Cham: Springer.
- Jamilly, M. 2019. *Limitations of AI*. BCS: The Chartered Institute for Information Technology. [https://www.bcs.org/content-hub/limitations-of-ai/?utm\\_source=british%20computer%20society&utm\\_medium=email&utm\\_campaign=10898395\\_25%2f09%20weekly%20newsletter%20-%20it%20now%20logo&utm\\_content=limitations%20of%20ai](https://www.bcs.org/content-hub/limitations-of-ai/?utm_source=british%20computer%20society&utm_medium=email&utm_campaign=10898395_25%2f09%20weekly%20newsletter%20-%20it%20now%20logo&utm_content=limitations%20of%20ai) [Accessed: 25 September 2019].
- Janet, B. & Reddy, A.V. 2010. Cube index: a text index model for retrieval and mining. *International Journal of Computer Applications*, 1(9):88-92.
- Jimmy, Zuccon, G. & Koopman, B. 2018. QUT IELab at CLEF 2018 consumer health search task: knowledge base retrieval for consumer health search. *Working Notes*. Conference and Labs of the Evaluation Forum (CLEF 2018), Avignon, France, 10-14 September.
- Joyce, J. 1932. *Ulysses*. 1932 ed. Ware: Wordsworth Classics.
- Kadous, M.W. 2002. Temporal classification: extending the classification paradigm to multivariate time series. Unpublished PhD thesis, University of New South Wales, Sydney.
- Kang, G., Liu, J., Tang, M., Cao, B. & Xu, Y. 2015. An effective web service ranking method via exploring user behavior. *IEEE Transactions on Network and Service Management*, 12(4):554-564.
- Keen, M. 1992a. Some aspects of proximity searching in text retrieval systems. *Journal of Information Science*, 18:89-98.
- Keen, M. 1992b. Term position ranking: Some new test results. *Proceedings: The 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, 21 - 24 June 1992: SIGIR92.

- Kelly, D. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1):1-224.
- Kent, A., Berry, M.M., Luehrs Jr, F.U. & Perry, J.W. 1955. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation*, 6(2):93-101.
- Khan, H., Caruso, B., Corson-Rikert, J., Dietrich, D., Lowe, B. & Steinhart, G. 2010. DataStaR: Using the semantic Web approach for data curation. *Proceedings: The 6th International Digital Curation Conference*, Chicago, USA, 6 - 8 December 2010.
- King, G., Keohane, R.O. & Verba, S. 1994. *Designing social inquiry: scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.
- Kissel, Z.A. & Wang, J. 2017. Generic adaptively secure searchable phrase encryption. *Proceedings on Privacy Enhancing Technologies*, 2017(1)1:4-20.
- Kleinberg, J.M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5).
- Kobayashi, V., Mol, S. & Kismihók, G. 2015. Discovering learning antecedents in learning analytics literature. Unpublished paper, University of Amsterdam, Amsterdam.
- Kohavi, R. & Provost, F. 1998. Glossary of terms. *Machine Learning*, 30(2/3):271-274.
- Koopman, B. 2014. Why assessing relevance in medical IR is demanding. *Proceedings. The 37<sup>th</sup> International ACM SIGIR Conference on Research & Development in Information Retrieval*, Queensland, Australia, 6-11 July: 16-19.
- Koopman, B., Russell, J. & Zuccon, G. 2018. Task-oriented search for evidence-based medicine. *International Journal on Digital Libraries*, 19(2-3):217-229.
- Koopman, B. & Zuccon, G. 2019. A full-day from consumers to clinicians. WSDM 2019 tutorial on health search. *Proceedings. The 12<sup>th</sup> ACM International Conference on Web Search and Data Mining (WSDM 2019)*, Melbourne, Australia, 11-15 February.
- Koopman, B., Zuccon, G., Bruza, P., Sitbon, L. & Lawley, M. 2016. Information retrieval as semantic inference: a graph inference model applied to medical search. *Information Retrieval*, 19:6-37.
- Korda, N. 2019. *Machine learning and the rise of the 'citizen data scientist'*. BCS: The Chartered Institute for Information Technology. [https://www.bcs.org/content-hub/machine-learning-and-the-rise-of-the-citizen-data-scientist/?utm\\_source=british%20computer%20society&utm\\_medium=email&utm\\_campaign=10581707\\_29%2f05%20weekly%20newsletter%20-%20it%20now%20logo&utm\\_content=ml%20text](https://www.bcs.org/content-hub/machine-learning-and-the-rise-of-the-citizen-data-scientist/?utm_source=british%20computer%20society&utm_medium=email&utm_campaign=10581707_29%2f05%20weekly%20newsletter%20-%20it%20now%20logo&utm_content=ml%20text) [Accessed: 29 May 2019].
- Korde, V. & Mahender, C.M. 2012. Text classification and classifiers: a survey. *International Journal of Artificial Intelligence & Applications*, 3(2):85-99.
- Kuhlthau, C.C. 1991. Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361-371.



- Lahiri, T., Reames, M.A., Edson, K., Goyal, N., Makino, K., Patthak, A., Thomas, D., Sarkar, C., Hoang, C. & Jiang, Q. 2019. *Combined row and columnar storage for in-memory databases for OLTP and analytics workloads*. Justia Patents. <https://patents.justia.com/patent/20190197026> [Accessed: 13 August 2019].
- Landis, J.R. & Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159-174.
- Lang, K. 1995. NewsWeeder: Learning to filter netnews. *Proceedings. The 12<sup>th</sup> International Conference on Machine Learning (ICML 1995)*, Tahoe City, CA, USA, 9-12 July: 331-339.
- Langville, A.N. & Meyer, C.D. 2007. Information retrieval and Web search. In Rosen, K.H. & Hogben, L. (eds.), *Discrete mathematics and its applications: handbook of linear algebra*. Boca Raton, FL: Chapman & Hall/CRC:63.1-63.16.
- Levenshtein, V.I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 164(4):845-848.
- Lewis, A.A. 2010. Enterprise users and Web search behavior. Unpublished MSc thesis, University of Tennessee, Knoxville, TN.
- Lewis, D.D., Yang, Y., Rose, T.G. & Li, F. 2004. RCV1: a new benchmark collection for text categorisation research. *Journal of Machine Learning Research*, 5:361-397.
- Liao, C., Tsao, Y., Lu, X. & Kawai, H. 2019. *Incorporating symbolic sequential modelling for speech enhancement*. ResearchGate. [https://www.researchgate.net/publication/332778941\\_incorporating\\_symbolic\\_sequential\\_modeling\\_for\\_speech\\_enhancement](https://www.researchgate.net/publication/332778941_incorporating_symbolic_sequential_modeling_for_speech_enhancement) [Accessed: 15 August 2019].
- Liu, M., Fang, Y., Choulos, A.G., Park, D.H. & Hu, X. 2017. Product review summarisation through question retrieval and diversification. *Information Retrieval Journal*, 20(6):575-605.
- Losee, R.M. 2006. Browsing mixed structured and unstructured data. *Information Processing and Management*, 42(2):440-452.
- Luhn, H.P. 1953. A new method of recording and searching information. *Journal of the Association for Information Science and Technology*, 4(1):14-16.
- Luhn, H.P. 1957. A statistical approach to mechanised encoding and searching of literary information. *IBM Journal*, 1(4):309-317.
- Maddalena, E., Mizzaro, S., Scholer, F. & Turpin, A. 2017. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems*, 35(3):19.1-19.32.
- Mandelbrot, B. 1953. An informational theory of the statistical structure of languages. In Jackson, W. (ed.), *Communication theory*. London: Butterworth: 486-502.
- Manning, C., Nayak, P. & Raghavan, P. 2017. *Introduction to information retrieval – CS276 information retrieval and Web search – efficient scoring*. Stanford University. [https://web.stanford.edu/class/cs276/handouts/efficient\\_scoring\\_cs276\\_2013\\_6.pdf](https://web.stanford.edu/class/cs276/handouts/efficient_scoring_cs276_2013_6.pdf) [Accessed: 11 October 2018].
- Manning, C.D., Raghavan, P. & Schütze, H. 2008. *An introduction to information retrieval*. New York, NY: Cambridge University Press.

- Manwar, A.B., Mahalle, H.S., Chinchkhede, K. & Chavan, V. 2012. A vector space model for information retrieval: a MATLAB approach. *Indian Journal of Computer Science and Engineering*, 3(2):222-229.
- Mao, J., Lu, K., Mu, X. & Li, G. 2015. Mining document, concept, and term associations for effective biomedical retrieval: introducing MeSH-enhanced retrieval models. *Information Retrieval*, 18(5):413-444.
- Marais, M.A. 2016. Social capital as a resource in the Village Operator model for rural broadband internet access and use. Unpublished PhD thesis, University of Pretoria, Pretoria.
- March, S.T. & Smith, G.F. 1995. Design and natural science research on information technology. *Decision Support Systems*, 15(4):251-266.
- Markey, N. 2009. *Tame the BeaST: The B to X of BibTEX*. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.169.9276> [Accessed: 20 June 2019].
- Marrero, M., Sánchez-Cuadrado, S., Urbano, J., Morato, J. & Moreira, J. 2010. Information retrieval systems adapted to the biomedical domain. *El Profesional de la Información*, 19(3):246-254.
- Matveeva, I. & Levow, G. 2007. Topic segmentation with hybrid document indexing. *Proceedings. The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague (EMNLP-CoNLL), Czech Republic, 28-30 June: 351-359.
- Maxwell, K.T. 2014. Term selection in information retrieval. Unpublished PhD thesis, University of Edinburgh, Edinburgh.
- McCarthy, J. 1980. Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1-2):27-39.
- McCray, A.T. 1998. The nature of lexical knowledge. *Methods of Information in Medicine*, 37(4-5):353-360.
- Mercier, C. 2019. *30 years ago the world changed forever*. W3C. <https://www.w3.org/blog/2019/03/30-years-ago-the-world-changed-forever/> [Accessed: 14 August 2019].
- Metzler, D. & Croft, W.B. 2005. A Markov random field model for term dependencies. *Proceedings. The 28<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, Salvador, Brazil, 15-19 August: 472-479.
- Min, J., Leveling, J., Zhou, D. & Jones, G.J.F. 2010. Document expansion for image retrieval. *Proceedings. The 9<sup>th</sup> RIAO Conference Adaptivity, Personalisation and Fusion of Heterogeneous Information (RIAO 2010)*, Paris, France, 28-30 April.
- Mitra, A. & Awekar, A. 2017. On low overlap among search results of academic search engines. Unpublished paper, Indian Institute of Technology, Assam.
- Mitra, B., Diaz, F. & Craswell, N. 2017. Learning to match using local and distributed representations of text for Web search. *Proceedings. The 26<sup>th</sup> International World Wide Web Conference (IW3C2)*, Perth, Australia, 3-7 April.

- Mooers, C.N. 1950. The theory of digital handling of non-numerical information and its implications to machine economics. *Proceedings*. The meeting of the Association for Computing Machinery at Rutgers University, New Jersey, USA, 28-29 March. Zator Co.
- Mooers, C.N. 1951. *Scientific information retrieval systems for machine operation – case studies in design*. eBook: Zator Co.
- Mouton, J. 2004. *How to succeed in your Master's and Doctoral studies*. Pretoria: Van Schaik Publishers.
- Muller, H. & Holzinger, A. 2019. Kandinsky patterns. Unpublished paper, Cornell University, Ithaca, NY.
- Myers, M.D. 2004. *Qualitative research in information systems*. Association for Information Systems. <http://www.qual.auckland.ac.nz> [Accessed: 8 August 2019].
- Myers, M.D. & Klein, H.K. 2011. A set of principles for conducting critical research in information systems. *MIS Quarterly*, 35(1):17-36.
- Narayan, V., Yadav, R.D.S., Mehta, R.K., Rai, M., Ahmed, M., Maurya, R., Kanujiya, A. & Dharpal. 2017. A novel approach for information retrieval using Web based search engine. *International Journal of Current Engineering and Technology*, 7(3):1214-1220.
- National Institute of Standards and Technology. 2014. *TREC Statement on product testing and advertising*. TREC. <http://trec.nist.gov/trec.disclaim.html> [Accessed: 5 August 2019].
- National Institute of Standards and Technology. 2018. *The 27<sup>th</sup> Text REtrieval Conference, TREC 2018 Proceedings*. NIST Special Publication: SP 500-331. <https://trec.nist.gov/pubs/trec27/trec2018.html> [Accessed: 14 August 2019].
- Navarro, G. & Baeza-Yates, R. 2001. A hybrid indexing method for approximate string matching. *Journal of Discrete Algorithms*, 1(1):1-135.
- Nguyen, G.H., Tamine, L., Soulier, L. & Souf, N. 2018. A tri-partite neural document language model for semantic information retrieval. *Proceedings*. The 15<sup>th</sup> Extended Semantic Web Conference (ESWC 2018), Heraklion, Greece, 3-7 June.
- NIST see National Institute of Standards and Technology.
- NTCIR see NII Testbeds and Community for information access Research.
- NII Testbeds and Community for information access Research. 2019. *Test Collections*. NTCIR. <http://research.nii.ac.jp/ntcir/data/data-en.html> [Accessed: 2 May 2019].
- Onal, K.D., Zhang, Y., Altingovde, I.S., Rahman, M.M., Karagoz, P., Braylan, A., Dang, B., Chang, H. & Kim, H. 2018. Neural information retrieval at the end of the early years. *Information Retrieval*, 21(2-3):111-182.
- Orkphol, K. & Yang, W. 2019. Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet*, 11(5):114.
- Orlikowski, W.J. & Baroudi, J.J. 1991. Studying information technology in organisations: research approaches and assumptions. *Information Systems Research*, 2(1):1-28.

- Pal, D., Mitra, M. & Bhattacharya, S. 2015. *Exploring query categorisation for query expansion: a study*. arXiv:1509.05567v1 [cs.IR]. Cornell University, Ithaca, NY.
- Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X. & Ward, R. 2016. *Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval*. arXiv:1502.06922v3 [cs.CL]. Cornell University, Ithaca, NY.
- Panigrahi, D. & Gollapudi, S. 2013. Document selection for tiered indexing in commerce search. *Proceedings*. The 6<sup>th</sup> ACM International Conference on Web Search and Data Mining (WSDM 2013), Rome, Italy, 4-8 February: 73-82.
- Partridge, C. 2008. The technical development of Internet email. *Annals of the History of Computing*, 30(2):3-29. IEEE.
- Pather, S. 2006. E-commerce information systems (ECIS) success: a South African study. Unpublished DTech thesis, Cape Peninsula University of Technology, Cape Town.
- Patil, A., Dave, K. & Varma, V. 2013. Leveraging latent concepts for retrieving relevant ads for short text. *Proceedings*. The 35<sup>th</sup> European Conference on Information Retrieval (ECIR 2013), Moscow, Russia, 24 - 27 March.
- Patil, H.J. & Surwade, Y.P. 2018. Web technologies from Web 2.0 to Web 4.0. *International Journal for Scientific Research & Development*, 4(4):810-814.
- Peppers, K., Tuunanen, T., Rothenberger, M.A. & Chatterjee, S. 2007. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45-78.
- Perry, J.W., Berry, M.M., Luehrs Jr, F.U. & Kent, A. 1954. Automation of information retrieval. *Proceedings*. The Eastern Joint Computer Conference: Design and Application of Small Digital Computers (AIEE-IRE 1954), Philadelphia, PA, USA, 8-10 December: 68-73.
- Pfeiffer, F., Broicher, A., Gillich, T., Klee, K., Mejia, J., Rampp, M. & Oesterhelt, D. 2008. Genome information management and integrated data analysis with HaloLex. *Archives of Microbiology*, 190(3):281-299.
- Pierce, C.S. 1958. *The collected papers of Charles Sanders Pierce (1931-1935)*. Cambridge, MA: Harvard University Press.
- Pinkerton, B. 2000. WebCrawler: finding what people want. Unpublished PhD thesis, University of Washington, Seattle, WA.
- Pinski, G. & Narin, F. 1976. Citation influence for journal aggregates of scientific publications: theory, with application to the literature of Physics. *Information Processing & Management*, 12(5):297-326.
- Popper, K. 1978. Three worlds: the tanner lecture on human values. Unpublished paper, University of Michigan, Ann Arbor, MI.
- Porter, M.F. 1980. An algorithm for suffix stripping. *Program*, 14(3):130-137.
- Powell, N. 2004. *Don't panic! Clear answers to your computer questions*. London: HarperCollins.

- Pratt, E.J. 1931. *Erosion*. University of Toronto Libraries.  
<http://canpoetry.library.utoronto.ca/pratt/poem1.htm> [Accessed: 14 August 2019].
- Procházka, P. & Holub, J. 2017. Towards efficient positional inverted index. *Algorithms*, 10(30):1-14.
- Rennie, D.M. & Jaakkola, T. 2005. Using term informativeness for named entity detection. *Proceedings*. The 28<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), Salvador, Brazil, 15-19 August: 353-360.
- Robertson, S. 2005. On event spaces and probabilistic models in information retrieval. *Information Retrieval*, 8(2):319-329.
- Robertson, S.E. 1981. The methodology of information retrieval experiment. In Spark Jones, K. (ed.), *Information retrieval experiments*. London: Butterworths: 9-31.
- Robson, C. 2005. *Real world research*. 2<sup>nd</sup> ed. Oxford: Blackwell.
- Rocchio, J.J. & Salton, G. 1965. Information search optimisation and iterative retrieval techniques. Unpublished paper, Harvard University, Cambridge, MA.
- Rocchio, J.J. 1965. Relevance feedback in information retrieval. Unpublished paper, Harvard University, Cambridge, MA.
- Rose, D.E. & Stevens, C. 1996. V-Twin: a lightweight engine for interactive use. *Proceedings*. The 5<sup>th</sup> Text REtrieval Conference (TREC-5), Gaithersburg, MD, USA, 20-22 November: 279-290.
- Rossi, C., De Moura, E.S., Carvalho, A.L. & Da Silva, A.S. 2013. Fast document-at-a-time query processing using two-tier indexes. *Proceedings*. The 36<sup>th</sup> International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013), Dublin, Ireland, 28 July-1 August.
- Ruthven, I. & Lalmas, M. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2):95-145.
- Ryan, A.B. 2006a. Methodology: collecting data. In Antonesa, M., Fallon, H., Ryan, A.B., Ryan, A., Walsh, T. & Borys, L. (eds.), *Researching and writing your thesis: a guide for postgraduate students*. Maynooth: Maynooth Adult and Community Education: 70-89.
- Ryan, A.B. 2006b. Post positivist approaches to research. In Antonesa, M., Fallon, H., Ryan, A.B., Ryan, A., Walsh, T. & Borys, L. *Researching and writing your thesis: a guide for postgraduate students*. Maynooth: Maynooth Adult and Community Education: 12-26.
- Sadiku, M.N.O., Shadare, A.E. & Musa, S.M. 2019. In-memory computing. *International Journal of Engineering Research and Advanced Technology*, 5(3):55-58.
- Salton, G. & Buckley, C. 1983. *Introduction to modern information*. New York, NY: McGraw-Hill.
- Salton, G. & Buckley, C. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513-523.
- Salton, G., Wong, A. & Yang, C.S. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 8(11):613-620.

- Sanderson, M. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247-375.
- Sankhavara, J. 2018. Biomedical document retrieval for clinical decision support system. *Proceedings*. The 56<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (ACL 2018), Melbourne, Australia, 15-20 July.
- Saunders, M., Lewis, P. & Thornhill, A. 2009. *Research methods for business students*. 5<sup>th</sup> ed. Harlow: Pearson Education.
- Saunders, M., Lewis, P. & Thornhill, A. 2016. *Research methods for business students*. 7<sup>th</sup> ed. Harlow: Pearson Education.
- Saunders, M., Lewis, P. & Thornhill, A. 2019. *Research methods for business students*. 8<sup>th</sup> ed. Harlow: Pearson Education.
- Scells, H., Zuccon, G. & Koopman, B. 2019. Automatic Boolean query refinement for systematic review literature search. *Proceedings*. The 2019 Web conference (TheWebConf 2019), San Francisco, CA, USA, 13-17 May.
- Scholer, F., Kelly, D. & Carterette, B. 2010. Information retrieval evaluation using test collections. *Information Retrieval Journal*, 19(3):225-229.
- Seddon, P.B., Staples, S., Patnayakuni, R. & Bowtell, M. 1999. Dimensions of information systems success. *Communications of the Association for Information Systems*, 2(20):1-61.
- Sedgwick, P. 2014. Unit of observation versus unit of analysis. *British Medical Journal*, 348:g3840.
- Shakespeare, W. 2018. *Hamlet*. 9<sup>th</sup> ed. Cape Town: Pearson.
- Shekarpour, S., Marx, E., Auer, S. & Sheth, A. 2017. RQUERY: Rewriting natural language queries on knowledge graphs to alleviate the vocabulary mismatch problem. *Proceedings*. The 31<sup>st</sup> AAAI Conference on Artificial Intelligence (AAAI 2017), San Francisco, CA, USA, 4-9 February.
- Shoaf, E.C. 2013. Cyril W. Cleverdon: his contributions to the theory of indexing and information retrieval. *Sci-Tech News*, 42(1):5-7.
- Simon, H.A. 1955. On a class of skew distribution functions. *Biometrika*, 42:425-440.
- Simon, H.A. 1996. *The sciences of the artificial*. 3<sup>rd</sup> ed. Cambridge, MA: MIT Press.
- Singhal, A. 2001. Modern information retrieval: a brief overview. *IEEE Data Engineering Bulletin*, 24(4):35-43.
- Sirres, R., Bissyandé, T.F., Kim, D., Lo, D., Klein, J. & Le Traon, Y. 2018. *Augmenting and structuring user queries to support efficient free-form code search*. Empirical Software Engineering. <https://doi.org/10.1007/s10664-017-9544-y> [Accessed: 5 February 2019].
- Skrlj, B., Martinc, M. & Pollak, S. 2019. tax2vec: constructing interpretable features from taxonomies for short text classification. Unpublished paper, Cornell University, Ithaca, NY.

- Smucker, M.D., Allan, J. & Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. *Proceedings*. The 16<sup>th</sup> ACM Conference on information and Knowledge Management (CIKM 2007), Lisbon, Portugal, 6-10 November: 623-632.
- Smuts, J.L. 2011. A knowledge management framework for information systems outsourcing. Unpublished PhD thesis, University of South Africa, Pretoria.
- Soldaini, L., Yates, A., Yom-Tov, E., Frieder, O. & Goharian, N. 2016. Enhancing web search in the medical domain via query clarification (Abstract Only). *Information Retrieval*, 19(1/2):149-173.
- Spärck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11-21.
- Sticht, T.G., Beck, L.J. & Hauke, R.N. 1974. *Auding and reading: a developmental model*. Alexandria, VA: Human Resources Research Organisation.
- Stokes, N., Li, Y., Cavedon, L. & Zobel, J. 2009. Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval*, 12(1):17-50.
- Takeda, H., Veerkamp, P., Tomiyama, T. & Yoshikawa, H. 1990. Modelling design processes. *AI Magazine*, 11(4):37-48.
- Tang, R. 1999. Use of relevance criteria across stages of document evaluation: a micro level and macro level analysis. Unpublished PhD thesis, University of North Carolina, Chapel Hill, NC.
- Taube, M. 1956. Machine retrieval of information. *Library Trends*, 5(2):301-308.
- Thuan, N.H., Drechsler, A. & Antunes, P. 2019. Construction of design science research questions. *Communications of the Association for Information Systems*, (forthcoming).
- Tijani, O.D., Akinwale, A.T., Onashoga, S.A. & Adeleke, E.O. 2017. An auto-generated approach of stop words using aggregated analysis. *Proceedings*. The 13<sup>th</sup> International Conference of the Nigeria Computer Society (NCS 2017), Abuja, Nigeria, 18-20 July: 99-115.
- Tolias, G. & Jégou, H. 2013. Local visual query expansion: exploiting an image collection to refine local descriptors. Unpublished report, National Technical University of Athens, Athens.
- Tonta, Y.A. 2019. *Failure analysis in document retrieval systems: a critical review of studies*. <http://yunus.hacettepe.edu.tr/~tonta/yayinlar/phd/bolum-3.htm> [Accessed: 23 April 2019].
- Tordai, A. 2006. Stem, stemming, stemmer: On the benefits of stemming in Hungarian. Unpublished Master's thesis, University of Amsterdam, Amsterdam.
- Transier, F. & Sanders, P. 2008. Out of the box phrase indexing. *Proceedings*. The 15<sup>th</sup> International Symposium on String Processing and Information Retrieval (SPIRE 2008), Melbourne, Australia, 10-12 November: 200-211.
- Trieschnigg, D. 2010. Proof of concept: concept-based biomedical information retrieval. Unpublished PhD thesis, University of Twente, Enschede.
- Trochim, W.M.K. 2006. *Social research methods*. The Research Methods Knowledge Base. <http://www.socialresearchmethods.net/kb/index.php> [Accessed: 4 April 2019].

- Troshin, P.V., Postis, V.L.G., Ashworth, D., Baldwin, S.A., McPherson, M.J. & Barton, G.J. 2011. PIMS sequencing extension: a laboratory information management system for DNA sequencing facilities. *BMC Research Notes*, 4:48.
- Tsikrika, T. & Lalmas, M. 2004. Combining evidence for Web retrieval using the inference network model: an experimental study. *Information Processing and Management*, 40(5):751-772.
- Turtle, H. & Croft, W.B. 1991. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187-222.
- Vaishnavi, V. & Kuechler, W. 2004. *Design science research in information systems*. DESRIST. <http://desrist.org/design-research-in-information-systems/> [Accessed: 28 January 2015].
- Vaishnavi, V., Kuechler, W. & Petter, S. 2019. Design science research in information systems. *DESRIST*. <http://www.desrist.org/design-research-in-information-systems/> [16 August 2019].
- Van den Bosch, A., Bogers, T. & De Kunder, M. 2016. Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics*, 106(2):839-856.
- Van Gysel, C., De Rijke, M. & Kanoulas, E. 2017. Neural vector spaces for unsupervised information retrieval. Unpublished paper, Cornell University, Ithaca, NY.
- Van Gysel, C., Li, D. & Kanoulas, E. 2018. *ILPS at TREC 2017 Common Core Track*. arXiv:1801.10603 [cs.IR], Cornell University, Ithaca, NY.
- Van Gysel, C.J.H. 2017. Remedies against the vocabulary gap in information retrieval. Unpublished PhD thesis, University of Amsterdam.
- Van Rijsbergen, C.J. 1979. *Information retrieval*. 2<sup>nd</sup> ed. London: Butterworths.
- Van Rijsbergen, C.J. 2004. *The geometry of information retrieval*. New York, NY: Cambridge University Press.
- Von Bertalanffy, L. 1968. *General systems theory*. New York, NY: George Braziller.
- Voorslys, W., Broberg, J. & Buyya, R. 2011. Introduction to cloud computing. In Buyya, R., Broberg, J. & Goscinski, A. (eds.), *Cloud computing: principles and paradigms*. New York: Wiley: 1-44.
- Waitelonis, J. 2018. Linked data supported information retrieval. Unpublished DEng thesis, Karlsruhe Institute of Technology, Karlsruhe.
- Wang, D. 2019. A Markov random field and adaptive regularisation embedded level set segmentation model solving by graph cuts. *Journal of Electrical and Computer Engineering*, 2019, Article ID 8747385. <https://doi.org/10.1155/2019/8747385>.
- Wang, Y., Huang, H. & Feng, C. 2017. Query expansion based on a feedback concept model for microblog retrieval. *Proceedings*. The 26<sup>th</sup> International Conference on World Wide Web (WWW 2017), Perth, Australia, 3-7 April: 559-568.



- Weideman, M. 2001. Internet searching as a study aid for information technology and information systems learners at a tertiary level. Unpublished PhD thesis, University of Cape Town, Cape Town.
- Wilkinson, R., Zobel, J. & Sacks-Davis, R. 1995. Similarity measures for short queries. *Proceedings. The 4<sup>th</sup> Text REtrieval Conference (TREC-4)*, Gaithersburg, MD, USA, 1-3 November: 277-285.
- Williams, H.E., Zobel, J. & Bahle, D. 2004. Fast phrase querying with combined indexes. *ACM Transactions on Information Systems*, 22(4):573-594.
- Wilson, T.D. 2000. Human information behavior. *Information Science*, 3(2):49-55.
- Yang, X., Hou, Y. & He, H. 2019. A processing-in-memory architecture programming paradigm for wireless Internet-of-Things applications. *Sensors*, 19(1):140.
- Yeasmin, S. & Rahman, K.F. 2012. 'Triangulation' research method as the tool of social science research. *BUP Journal*, 1(1):154-163.
- Yu, B. 2019. Research on information retrieval model based on ontology. *EURASIP Journal on Wireless Communications and Networking*, Volume 2019, Article No. 30. SpringerOpen.
- Zamperi, F.A., Adli, N.H.H., Hussin, N. & Ahmad, M. 2018. Information retrieval via social media. *International Journal of Academic Research in Business and Social Sciences*, 8(12):1375-1381.
- Zhao, L. & Huang, H. 2016. Application of related data automatic semantic annotation technology in Internet of things. *Proceedings. The 2<sup>nd</sup> Workshop on Advanced Research and Technology in Industry Applications (WARTIA 2016)*, Dalian, China, 14-15 May: 1262-1266.
- Zhao, L. 2012. Modelling and solving term mismatch for full-text retrieval. Unpublished PhD thesis, Carnegie Mellon University, Pittsburgh, PA.
- Zipf, G.K. 1935. *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin.
- Zipf, G.K. 1949. *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.
- Zipf, G.K. 1965. *The psycho-biology of language: an introduction to dynamic philology*. Cambridge: MIT Press.