http://sajlis.journals.ac.za doi:10.7553/ 10.7553/87-2-1957

Information retrieval: Solving mismatching vocabulary in closed document collections

Kyle Andrew Fitzgerald[1], Andre Charles de la Harpe[2], Corrie Susanna Uys[3] and Andrew John Bytheway[4]

kyle@chekadee.co.za ORCID: 0000-0003-0325-5386
delaharpea@cput.ac.za ORCID: 0000-0003-0100-556X
uysc@cput.ac.za ORCID: 0000-0002-0313-3875
andy.bytheway@gmail.com ORCID: 0000-0003-1603-0581

Received: 16 August 2020
Accepted: 9 December 2021

During a search, phrase-terms expressed in queries are presented to an information retrieval system (IRS) to find documents relevant to a topic. The IRS makes relevance judgements by attempting to match vocabulary in queries to documents. If there is a mismatch, the problem of vocabulary mismatch occurs. The aim is to examine ways of searching for documents more effectively, in order to minimise mismatches. A further aim is to understand the mechanisms of, and the differences between, human and machine-assisted, retrieval. The objective of this study was to determine whether IRS-H (an IRS using the hybrid indexing method) and human participants agree or disagree on relevancy judgments, and whether the problem of mismatching vocabulary can be solved. A collection of eighty research documents and sixty-five phrase-terms were presented to (i) IRS-H and four participants in Test 1, and (ii) IRS-H and one participant (aided by search software) in Test 2. Statistical analysis was performed using the Kappa coefficient. IRS-H and the four participants' judgements disagreed. IRS-H and the participant aided by search software judgments did agree. IRS-H solves the problem of mismatching vocabulary between a query and a document.

Keywords: Information retrieval, hybrid indexing method, precision, recall, research, vocabulary mismatch

1 Introduction

During mid-2019 one of the authors of this research visited the British Library in London and had an engaging conversation with a doctor – a general practitioner with an interest in medical research. The conversation quickly focused on the problem of searching large, closed collections of research papers. Over his career, the doctor's collection had grown substantially, covering many research disciplines, and he bewailed the difficulty in finding documents relevant to his needs. This anecdotal evidence indicates a problem that is shared by many, as our personal as well as institutional collections become larger and more challenging to manage. Koopman et al. (2016), who agree that there exists this information retrieval problem, described a model for complex queries with the aim of improving information retrieval in the biomedical domain, while He and Ounis (2009) investigated query expansion effectiveness in an attempt to improve the retrieval of information.

The aim of this research was to find a solution to the problem of a mismatch in vocabulary between words used in a query and words used in a document (Koopman et al. 2016). The research was performed by determining whether an Information Retrieval System (IRS) using the hybrid indexing method could solve the problem through: (i) examining ways of searching for documents more effectively to minimise mismatches; and (ii) understanding the mechanisms that play a role in, and the differences between, human and machine-assisted retrieval (Bauer et al. 2009). This research therefore investigated typical mechanisms used in: (i) retrieving the desired documents from a collection, for example: search queries, words used as terms within these queries (Manning, Raghavan & Schütze 2008); and (ii) the design and functionality of indices (Fitzgerald, Fitzgerald & Bytheway 2017) as well as how to improve on these to retrieve documents more effectively.

[1] Kyle Andrew Fitzgerald received his Doctor of Information and Communication Technology degree from the Faculty of Informatics and Design, Cape Peninsula University of Technology, South Africa

[2] Andre Charles de la Harpe is Senior Lecturer/Supervisor, Faculty of Informatics and Design, Cape Peninsula University of Technology, South Africa

[3] Corrie Susanna Uys is Statistical Consultant, Centre for Postgraduate Studies, Cape Peninsula University of Technology, South Africa

[4] Andrew John Bytheway is Old Mutual Chair in Information Systems (retired), University of the Western Cape, South Africa

1.1 Vocabulary mismatch

As experienced by the general practitioner, when using search mechanisms to search for documents, the problem of vocabulary mismatch may occur. Vocabulary mismatch is a phenomenon which is twofold. The vocabulary mismatch problem relates to (i) how words are presented as phrases and then expressed within queries, and (ii) how the index stores and handles the text acquired from the document and then communicates with the query (Shekarpour et al. 2017, Nguyen et al. 2018, Onal et al. 2018).

To search for documents effectively, the user firstly needs to choose the words for the specific search carefully and must express these words in the query in the correct format. The ordering of the words is important and so is the proximity of words. The vocabulary mismatch problem is compounded when using multi-word phrases, referred to as 'phrase-terms', rather than singular words. For example, if the user wishes to find documents that contain the multi-word phrase ‘type 1 diabetes mellitus’ (Rother & Harlan 2004), the user expects a document to be retrieved from a closed document collection where that phrase occurred at least once. Synonymic phrases occur too when things are named and then renamed over time and this phenomenon complicates searching. The user must be well informed about the topic being explored and be aware of any synonymic phrases that exist. A classic example is in healthcare where naming conventions for diseases have changed, which creates difficulty – ‘type 1 diabetes mellitus’ was known in the past as ‘thin diabetes’ (Gale 2001), ‘juvenile diabetes’ (Jun & Yoon 2002), and ‘insulin dependent diabetes mellitus’ (Katahira 2009), and more recently has split into two different disease classifications: (i) ‘type 1a diabetes mellitus’ where an autoimmune process exists, and (ii) ‘type 1b diabetes mellitus’ where the cause is unknown (Gale 2006).

The way in which the words, acquired from a document, are stored in an index is also important. Langville and Meyer (2007) state that the *precision* of an IRS is the ratio of the number of relevant documents retrieved to the number of documents retrieved. The *recall* of an IRS is the ratio of the number of relevant documents retrieved to the number of relevant documents in a collection. Therefore, to be able to measure precision and recall the words: (i) must be linked to the documents they exist in, and (ii) must be stored in a way so that the phrase-term within the query can identify it. The index must present the words to the query in the correct word order (Clarke, Cormack & Tudhope 2000) and within the correct word proximity (Mitra, Diaz & Craswell 2017). The goal is not to keep a list of all forms of phrases but rather a list of words with unique (per document) token (word) identifiers. This allows each word acquired from the text to be stored in the index, and because of its own unique token identifier, the position of the word preceding it and succeeding it is known.

In reviewing the work of authors describing the vocabulary mismatch problem, it is found that the phrases themselves used to describe vocabulary mismatch create mismatching vocabulary problems, for example: ‘vocabulary problem’ (Furnas et al. 1987, Turtle & Croft 1991, Egozi, Markovitch & Gabrilovich 2000, Min et al. 2010, Shekarpour et al. 2017): ‘vocabulary gap’ (IJzereef, Kamps & De Rijke 2005, Liu et al. 2017, Van Gysel, De Rijke & Kanoulas 2017), ‘term mismatch’ (Zhao & Callan 2010, Sirres et al. 2018) and ‘semantic gap’ (Nguyen et al. 2018, Koopman & Zuccon 2019). Antonyms have been used to describe the opposite: ‘vocabulary normalisation’ (Binkley, Lawrie & Uehlinger 2012, Binkley & Lawrie 2015) and ‘vocabulary agreement’ (Chaparro, Florez & Marcus 2016). To perform a solid search and to retrieve relevant documents for the problem of vocabulary mismatch, all these synonymic and antonymic phrases, and possibly others, should be expressed within a query. These examples illustrate the complexity of the vocabulary mismatch problem and the burden on a researcher who must strive to become completely informed about a topic before performing a search.

1.2 Indexing mechanisms

In a literature review conducted on information retrieval methods, the researchers found a number of mechanisms and indices for retrieval, each having its own unique design properties. The concept of text indexing is not new. Gross and Gross (1927) described text indexing based on physically matching specific words within a language to those words found in the text of a document. A range of methods was found in later work. The often-cited ‘inverted index’ (Croft, Metzler & Strohman 2015) stores singular tokens (chunks of text in the form of words, acronyms and codes) acquired from the text of a document together with the document numbers in which they were found. When a term, expressed within a search query, is presented to the inverted index, the index attempts to match the term to a token stored within the index. If a match is found, the index returns those document numbers matched to the token stored within the index. If a match is not found the document number is omitted. However, this inverted index can only handle queries using single word terms. A word could exist on the first page and another on the last page of the document. These words without order are effectively treated as a “bag of words” (Harris 1954:156), where word ordinality and word proximity are lost. The literature review continued seeking not only a mechanism that could handle multi-word phrase queries, for example ‘type 1a diabetes mellitus’ (Gale 2006:3) but also an index that could maintain word ordinality and word proximity (Fitzgerald, Fitzgerald & Bytheway 2017). Some candidates that were found were:

- The ‘phrase query’ (Ha et al. 2002) handles the presentation of single word, bi-word, tri-word and multi-word queries but it does not store these phrases within an index.
- The ‘tiered index’ (Panigrahi & Gollapudi 2013, Manning, Nayak & Raghavan 2017) is an inverted index broken up into tiers of decreasing importance. Again, this index cannot handle phrases and maintain word ordinality and proximity.
- The ‘next word index’ (Williams, Zobel & Bahle 2004, Muller & Holzinger 2019) uses concepts from the phrase query and positional index, but it does not match the criteria required for this research.
- Fitzgerald, Fitzgerald and Bytheway (2017:2) introduced the hybrid indexing method that utilises a pair of indices, namely a ‘hybrid query index’ and a ‘hybrid token index’. Phrase-terms are presented within queries to the query index, and tokens acquired from document text are stored with unique token identity numbers within the token index. When a phrase is presented as a query, the two indices interrogate each other and attempt to perform a match. As word ordinality and word proximity is maintained, if an exact match occurs, the document number is returned from the token index in which that phrase occurs.

In this research, an IRS that uses the hybrid indexing method is used in an attempt to solve the problem of mismatching vocabulary experienced by the general practitioner.

1.3 An information retrieval system using the hybrid indexing method

When searching for documents within a closed document collection pertaining to a specific topic, multi-word phrase-terms in the form of strings are formulated by the human in an attempt to describe the criteria with which to search the topic. These phrase-terms are then expressed within a search query and presented to an IRS in the hope of finding documents relevant to the topic (Manning, Raghavan & Schütze 2008, Croft, Metzler & Strohman 2015). Based on the query, the search engine presents the query to the IRS’s index that contains the tokens acquired from the text within a document, and a reference is made to the document from which the token originated. These tokens are chunks of text (Bytheway 2014) that are either words (used within a language) or codes, special characters, abbreviations, acronyms (Bell et al. 1993) or other forms of novel text (Joyce 1932). Indexing design can vary, thereby affecting the IRSs precision and recall measurements, the judgments that are made by the IRS as to whether or not to retrieve the document as relevant (Williams, Zobel & Bahle 2004, Transier & Sanders 2008, Wang, Huang & Feng 2017).

An IRS using the hybrid indexing method (IRS-H), introduced in 2017 (Fitzgerald, Fitzgerald & Bytheway 2017), with an original design approach based on design science research (Hevner et al. 2004, Gregor & Hevner 2013, Hevner, Vom Brocke & Maedche 2019), was used in this study. The study uses a hybrid indexing method based on a pair of indices, for the query index and for the token index. The IRS-H method attempts to match a phrase-term expressed within a query within its query index to a sequentially ordered set of tokens stored within its token index. When a match is suggested by the indices, IRS-H indicates that, by its judgement, the query matches the document and therefore retrieves the document from the collection. Whereas in the past it was just a guess (Van Rijsbergen 1979), IRS-H attempts to match exactly one or more multi-word phrase-terms within a query to the phrases that exist within the text of the document. To measure how effectively an IRS has performed, human involvement is needed. The human is concerned with the ‘truth’ (Manning, Raghavan & Schütze 2008, Croft, Metzler & Strohman 2015) and needs to now perform a similar exercise to judge whether one or more phrase-terms expressed within a query actually exist within a document, so that the document can be judged by the human as relevant. However, the results of the judgements made by the IRS and the results of the judgements made by the human can differ. When this occurs, the IRS is often blamed for the difference in judgements (Croft 2019).

2 Hypotheses

Two null hypotheses guided this research:

- $H1_0$: Relevancy judgments made by humans unaided by search software and an IRS using the hybrid indexing method disagree
- $H2_0$: Relevancy judgments made by a human aided by search software and an IRS using the hybrid indexing method disagree

3 Materials and methods

The research adopted an objectivist view and a positivist stance, as a structured, quantitative methodology is used to enable the replication of research (Burrell & Morgan 1979) and because an IRS can use large volumes of records. Deductive reasoning, which is aligned to quantitative methodology, was found to be an appropriate research approach because the research used hypotheses, binary coding and empirical generalisations (Babbie 2013). The research method was cross-sectional as the data collected was a snapshot in time (Saunders, Lewis & Thornhill 2019). The research strategy followed was experimental (Tsikrika & Lalmas 2004). The populations of this research comprised (i) all academic literature available at Cape Peninsula University of Technology (CPUT) Libraries and (ii) Information Technology postgraduate students at CPUT. These populations were selected because: (i) the library was available, and (ii) the first author was a registered doctoral student at CPUT. The library provided a wide range of academic research material. The postgraduates were purposively selected from the Information Technology department because of their solid understanding of concepts relevant to this research.

A controlled environment was required to ensure completion of the experiment within the allocated time frame. As time for the experiment was limited to eight hours (one working day), four postgraduates were selected as participants and twenty documents were allocated to each participant which would allow approximately twenty-five minutes for each participant to read each of their twenty documents. Thus, a sample of eighty documents was randomly selected from published and unpublished academic literature sourced from CPUT Libraries, spanning a range of disciplines. A file listing browser was used to select these documents. This collection of eighty academic documents (N = 80) in the form of journal articles, conference papers, books and theses, was made available by the library for experimentation. A set of sixty-five phrase-terms, each relating to a single query, was compiled from the literature (Table 1).

Table 1 The 65 phrase-terms and queries used in this study

pt no	q no	Phrase-term/Query	pt no	q no	Phrase-term/Query	pt no	q no	Phrase-term/Query
pt01	q01	design science	pt23	q23	clinical guideline	pt45	q45	design research methods
pt02	q02	design sciences	pt24	q24	clinical guidelines	pt46	q46	design research method
pt03	q03	design science research	pt25	q25	clinical guidelines in primary care	pt47	q47	design research philosophy
pt04	q04	design science methodology	pt26	q26	clinical guidelines in family practice	pt48	q48	design research pragmatism
pt05	q05	the design method	pt27	q27	clinical guidelines for operations	pt49	q49	design theory
pt06	q06	design research	pt28	q28	clinical guidelines for stroke management	pt50	q50	data quality
pt07	q07	design science research paradigm	pt29	q29	cloud computing	pt51	q51	data qualities
pt08	q08	design science research paradigms	pt30	q30	cloud computing types	pt52	q52	data quality methodology
pt09	q09	qualitative method	pt31	q31	cloud computing models	pt53	q53	data quality methodologies
pt10	q10	qualitative analysis	pt32	q32	cloud computing service models	pt54	q54	data quality model
pt11	q11	qualitative research	pt33	q33	conceptual framework	pt55	q55	data quality models
pt12	q12	qualitative research design	pt34	q34	conceptual frameworks	pt56	q56	data quality conceptual models
pt13	q13	qualitative research method	pt35	q35	conceptual framework in research	pt57	q57	data quality conceptual model
pt14	q14	qualitative research methods	pt36	q36	conceptual frameworks in research	pt58	q58	data quality framework
pt15	q15	qualitative research methodology	pt37	q37	conceptual model	pt59	q59	data quality frameworks
pt16	q16	quantitative method	pt38	q38	conceptual models	pt60	q60	electronic health record
pt17	q17	quantitative analysis	pt39	q39	research ethics	pt61	q61	electronic health records
pt18	q18	quantitative research	pt40	q40	ethics in research	pt62	q62	e health record
pt19	q19	quantitative research design	pt41	q41	research ethics principles	pt63	q63	e health records
pt20	q20	quantitative research method	pt42	q42	design method	pt64	q64	electronic patient record
pt21	q21	quantitative research methods	pt43	q43	design methods	pt65	q65	electronic patient records
pt22	q22	quantitative research methodology	pt44	q44	design practice			

Three experiments using the IRS-H method were performed to generate the necessary data. The four participants were asked to judge document relevancy after each query. Two tests were then carried out to ascertain document relevancy judgements made by IRS-H and the human participants. Experiment 1 determined the judgements the four participants made, unaided by any software, when deciding whether a query was relevant or non-relevant to a document. Experiment 2 determined the judgements IRS-H made when deciding whether a query was relevant or non-relevant to a document. Experiment 3 determined the judgements the participants made, aided by search software (Adobe Reader XI with its Advanced Search feature), when deciding whether a query was relevant or non-relevant to a document.

The results obtained from the two tests:

- Test 1 compared those judgements made by the four participants in Experiment 1 to the judgements made by IRS-H in Experiment 2. The results from Test 1 were used to test the first of the two hypotheses.
- Test 2 compared those judgements made by IRS-H in Experiment 2 to the judgements made by the participant who was aided by search software in Experiment 3. The results from Test 2 were used to test the second hypothesis.

Data were analysed by performing a statistical analysis on the two sets of generated test data in order to test the two hypotheses.

3.1 Presenting the phrase-terms and queries

From the literature, sixty-five phrase-terms (pt) that described ten categories based on general research information needs were compiled. In an attempt to overcome the vocabulary mismatch problem, each category contained a number of synonymic phrase-terms that described that category in some way. Many more could exist, but these sixty-five phrase-terms had a mix of singular and plural word formats, and words used in British English and not United States English. These phrase-terms were all converted to lowercase (Ruthven & Lalmas 2003, Agnihotri, Verma & Tripathi 2017) and hyphenation (Markey 2009, Waitelonis 2018), suffix stripping (Porter 1980, Markey 2009), stemming (Frej, Chevallet & Schwab 2018) and special characters were omitted to suit the query input requirements of IRS-H. Note that each query (q) is represented by a single phrase-term representing a one-to-one relationship. Following the approach used by Fitzgerald, Fitzgerald and Bytheway (2017), based on the inverted index method and the hybrid index method, two to five-word phrase terms were selected. The retrieval process is based on the matching of the words in the correct order (Clarke, Cormack & Tudhope 2000) and within the same proximity (Mitra, Diaz & Craswell 2017). The phrase-term numbers, query numbers and their strings of text representing the phrase-terms and queries are presented in Table 1.

3.2 Experiment 1

The first experiment (Figure 1) measured human judgements, where the group of four postgraduate researchers participated in reading the twenty unique documents. After reading the documents, each participant completed a questionnaire. The questionnaire contained ten pages. Each page represented a category of mismatch vocabulary compiled using synonymic phrases where each phrase represented one of the sixty-five phrase-terms. For example, for the category of ‘design science research’, the question was:

> 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 a cross (false) whether each phrase term, pt01 through to pt08 (columns 2 to 9), exists within each of the documents.

The participants were required to indicate whether any of the phrase-terms specified in the questionnaire occurred in any of the documents provided. Answers to the questions were Boolean values of true or false and were stored in a phrase-term-by-document matrix (Kobayashi, Mol & Kismihók 2015). The Boolean values were then converted to binary values, where ‘true’ was set to ‘1’ and ‘false’ set to ‘0’ and were stored in a query-by-document matrix. If a cell within this matrix contained a ‘1’, this indicated the document was judged relevant by the human. Therefore, data generated by the questionnaire provided the number of documents judged relevant by the participants (‘1’) represented as the sum of true positive and false negative (tpfn) and the number of documents judged non-relevant by the participants (‘0’) as the sum of false positive and true negative (fptn) (Kohavi & Provost 1998), for each of the sixty-five questions.

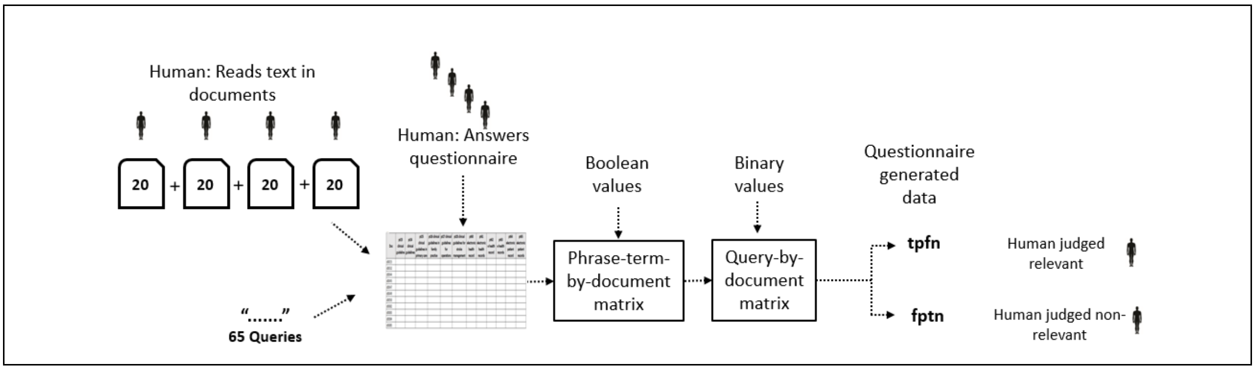

Human: Reads text in documents
20
+
20
+
20
+
20
"......"
65 Queries
Human: Answers questionnaire
Boolean values
Phrase-term-by-document matrix
Binary values
Query-by-document matrix
Questionnaire generated data
tpfn
fptn
Human judged relevant
Human judged non-relevant

Figure 1 Experiment 1

3.3 Experiment 2

The second experiment (Figure 2) measured IRS judgments. The way that the design mechanism of IRS-H and its hybrid indexing method differs from others is that this indexing method enables the IRS to make exact matches. This method maintains word ordinality and proximity; it matches phrase-terms in queries exactly to those phrase-terms that exist in the text of documents. All eighty documents in the collection were provided and the sixty-five phrase-terms (pt) presented to the IRS for evaluation. When the search is activated within IRS-H, the system generates phrase-term frequency (ptf) values – the number of times a phrase-term occurs in a document (Fitzgerald, Fitzgerald & Bytheway 2017) and stores these ptfs in a phrase-term-by-document matrix. The ptf values are then converted to binary values where '1' represents ptf > 0 and '0' represents ptf = 0 and are stored in a query-by-document matrix. If a cell within this matrix contains a '1', this indicates the document was judged relevant by IRS-H. The number of documents judged relevant by IRS-H ('1') is generated in the form of tpfp (the sum of true positives and false positives) and the number of documents judged non-relevant by IRS-H ('0') as fntn (the sum of false negatives and true negatives), for each of the sixty-five queries.

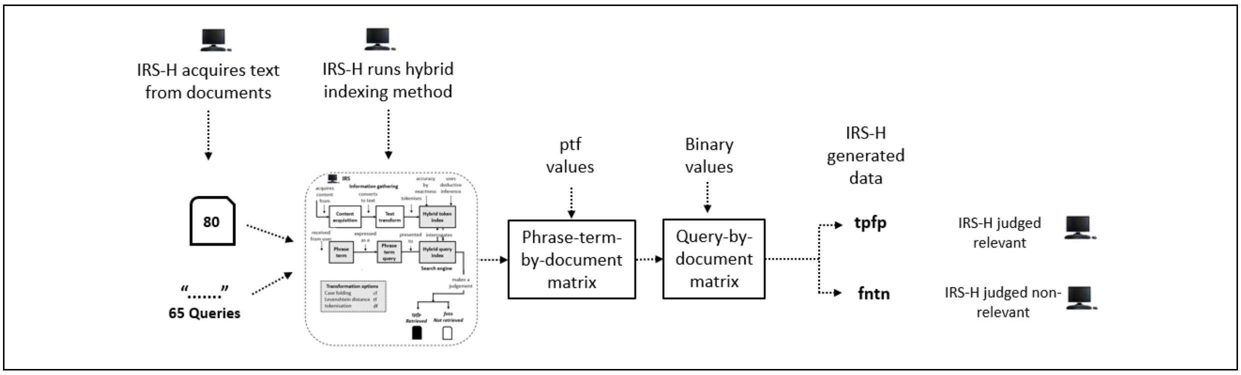

IRS-H acquires text from documents
IRS-H runs hybrid indexing method
80
"......"
65 Queries
ptf values
Phrase-term-by-document matrix
Binary values
Query-by-document matrix
IRS-H generated data
tpfp
fntn
IRS-H judged relevant
IRS-H judged non-relevant

Figure 2 Experiment 2

The design of IRS-H (Figure 3) ensures that, when searching for a document using multiple words in a phrase-term, expressed within a query, the ordinality and the proximity of the words are maintained. The design of IRS-H utilising the hybrid indexing method contains three artefacts, namely the IRS itself, and a pair of hybrid indices. The first index, the hybrid token index, is populated with the tokens acquired from the text of the documents within the collection. The second index, the hybrid query index, is populated with the phrase-terms expressed within the queries. When attempting to match a query to a document, the hybrid query index interrogates the hybrid token index and returns a result. By design, what IRS-H searches for is an exact match of a string, which accommodates multiple words, but not other combinations of those words at different proximities (Fitzgerald, Fitzgerald & Bytheway 2017).

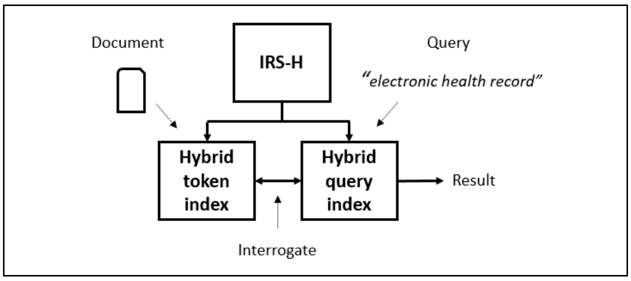


Figure 3 The indices of the hybrid indexing method

3.4 Experiment 3

The third experiment (Figure 4) repeated Experiment 1 but measured the judgements of a single participant when aided by search software. The search software was Adobe Reader XI using the functionality of its Advanced Search feature. One of the authors of this study participated in answering the same questionnaire used during the first experiment. This participating researcher was required to read all eighty documents in the collection, and to search for the specific phrase-terms using the search software. When the results were returned by the search software, the researcher counted the number of times a phrase-term occurred in a document, thus providing the value for ptf, and thereafter answered the questionnaire using the ptf value. The ptf values were then converted to binary values where ‘1’ represents ptf > 0 and ‘0’ represents ptf = 0 and were stored in a query-by-document matrix. If a cell within this matrix contained a ‘1’, this indicated the document was judged relevant by the human. The data generated from the questionnaire, answered by the human aided by the search software, provided the number of documents judged relevant by the human (‘1’) represented as the sum values of true positive and false negative (tpfn), and the number of documents judged non-relevant by the human (‘0’) represented as the sum values of false positive and true negative (fptn), for each of the sixty-five questions.

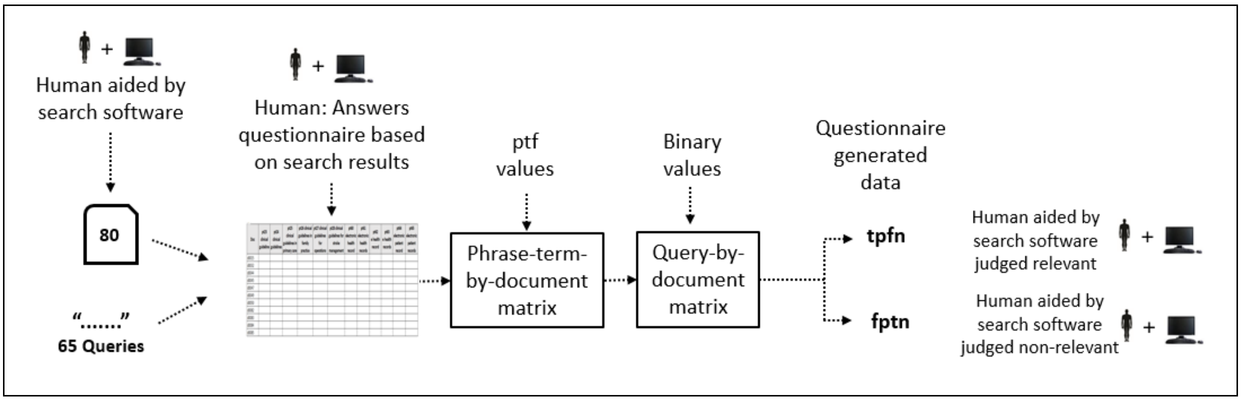

Human aided by search software
Human: Answers questionnaire based on search results
ptf values
Binary values
Questionnaire generated data
80
“......”
65 Queries
Phrase-term-by-document matrix
Query-by-document matrix
tpfn
fptn
Human aided by search software judged relevant
Human aided by search software judged non-relevant

Figure 4 Experiment 3

3.5 Test 1: IRS-H versus human judgements

The first test (Figure 5) was to test judgements made by IRS-H against those judgements made by the humans and to determine whether these judgements agreed or disagreed with each other. Using the sixty-five queries and the eighty documents in the collection, the participants were set as the control group and therefore those judgements made by the participants were considered to be the truth. IRS-H was set as the variable and those judgements made by IRS-H were then compared with those made by the participants. The results generated were processed through a 2x2 contingency table (Cleverdon 1967) and then analysed. The first null hypothesis was then tested to determine whether relevancy judgments made by the participant unaided by the search software and an IRS using the hybrid indexing method agreed or disagreed with each other.

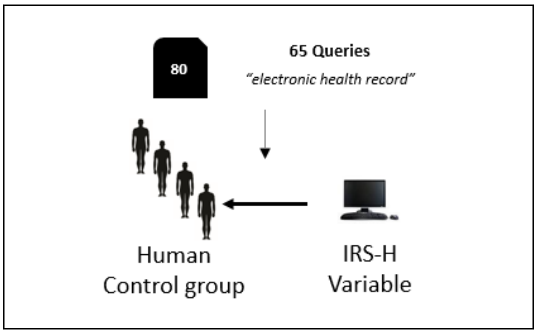


Figure 5 IRS-H versus human judgements

3.6 Test 2: Human judgements aided by search software versus IRS-H

The second test (Figure 6) was to test judgements made by one participant aided by search software against judgement made by IRS-H and to determine whether these judgements agreed or disagreed with each other. Using the sixty-five queries and the eighty documents in the collection, IRS-H was set as the control group and therefore those judgements made by IRS-H were considered to be the truth. The participant was set as the variable and those judgements made by the participant were then compared with those made by IRS-H. The results generated were processed through a 2x2 contingency table and then analysed. Thereafter, the second null hypothesis was tested to determine whether relevancy judgments made by a participant aided by search software and an IRS using the hybrid indexing method agreed or disagreed with each other.

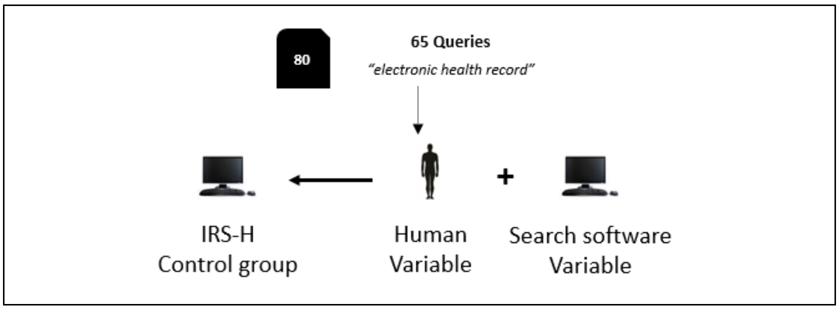


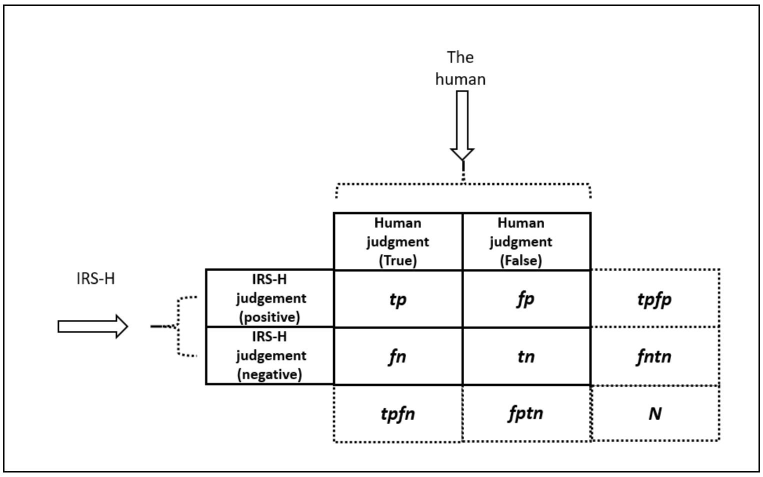
Figure 6 Human judgements aided by search software versus IRS-H

4 Data analysis

A 2x2 contingency table was used to collect the data in the format required (Figure 7). Reading from left-to-right, the top row represents the number of documents judged relevant (positive) made by IRS-H (tpfp) where the binary values from the query-by-document matrix are set at '1'. The bottom row represents the number of documents judged non-relevant (negative) by IRS-H (fntn) where the binary values are set at '0'. Reading from top-down, the first column represents the number of documents judged relevant (true) by the participants (tpfn) where the binary values from the query-by-document matrix are set at '1'. The second column represents the number of documents judged non-relevant (false) by the participants (fptn) where the binary values are set at '0'. Using the generated data from Experiments 1 and 2, the results acquired from the query-by-document matrices were placed into 2x2 contingency tables to determine the values of tpfp and fntn for IRS-H and the values of tpfn and fptn for the participants, for Test 1. The values for tp, fp, fn and tn were then derived from tpfp, fntn, tpfn and fptn where:

- tp: true positive – the number of participants judged relevant documents also judged relevant by IRS-H
- fp: false positive – the number of participant non-relevant documents, judged relevant by IRS-H
- tn: true negative – the number of participants judged non-relevant documents, also not judged relevant by IRS-H, and
- fn: false negative – the number of participants judged relevant documents, not judged relevant by IRS-H

Similarly, the data generated from Experiments 2 and 3 were processed and the values for tpfp, fntn, tpfn and fptn were calculated for Test 2. N refers to the total number of documents within the collection, thus N = 80.



	Human judgment (True)	Human judgment (False)	
IRS-H judgement (positive)	*tp*	*fp*	*tpfp*
IRS-H judgement (negative)	*fn*	*tn*	*fntn*
	tpfn	*fptn*	*N*

Figure 7 A 2x2 contingency table

IBM SPSS Statistics version 25 (SPSS) data analysis was performed using the results produced from the two tests. To support testing the hypotheses, the Kappa coefficient (k) (Cohen 1960, De Raadt et al. 2019) was used to measure the difference in judgements between the four participants and IRS-H in Test 1, and between IRS-H and the participant aided by search software in Test 2. The formula used was: k = (P(A) - P(E) / (1 - P(E)), where P(A) = (tp + tn) / N, P(E) = P(non-relevant)^2 + P(relevant)^2, P(non-relevant) = ((fn + tn) + (fp + tn))/(N + N), P(relevant) = ((tp + fp) + (tp + fn))/(N + N), and N = tp + fp + fn + tn. A six-division range of agreement measurements (Landis & Koch 1977, Fleiss, Levin & Paik 2003) was used to represent the results in order to assist in explaining the test results of the hypotheses and achieving the aim of this research. This six-division measurement range of Landis and Koch (1977:7) for the Kappa coefficient was presented as: k < 0.00 as poor, 0.00 ≤ k ≤ 0.20 as slight, 0.21 ≤ k ≤ 0.40 as fair, 0.41 ≤ k ≤ 0.60 as moderate, 0.61 ≤ k ≤ 0.80 as substantial, and 0.81 ≤ k ≤ 1.00 as almost perfect.

5 Results

The test results from the two tests, that used the data from the three experiments, are now presented. The first test, Test 1: IRS-H versus human judgements, used data generated by the four participants in answering the questionnaire in Experiment 1 and IRS-H data from Experiment 2. The data were then statistically analysed. The results are presented in Table 2. The second test, Test 2: Human judgements aided by search software versus IRS-H, used data generated by IRS-H from Experiment 2 and data generated by the single participant in answering the questionnaire from Experiment 3. The data were then statistically analysed. The results are presented in Table 3.

5.1 Test 1: IRS-H versus human judgements

Test 1 determined whether the judgments made by IRS-H and the humans agree, thus testing the first hypothesis. Table 2 presents the results of the 5,200 judgement cases analysed testing sixty-five queries in eighty documents.

Table 2 IRS-H versus human judgements

Judgement test	Query structure	No of docs	No of cases	Control group	Kappa coefficient (*k*)	Strength of agreement	Significance (*p*)	Statistically significant
IRS-H * Human	65 single queries each containing a single phrase-term	80	5200	Human	0.516	Moderate	p < 0.001	Yes

Test 1 was used to test *$H1_0$* where the strength of agreement between the judgements made by IRS-H in Experiment 2 and those judgements made by the participants in Experiment 1 were assessed. The results revealed a Kappa coefficient of k = 0.516 with a statistical significance of p < 0.001 and a strength of agreement of 'Moderate'. The results from the test show that the agreement in judgements between IRS-H and the participants was moderate, as in nearly one half of the cases agreements were made while in the other half, agreements were not made. $H1_0$ was therefore accepted, namely that *the relevancy judgments made by humans unaided by search software and an IRS using the hybrid indexing method disagree.*

5.2 Test 2: Human judgements aided by search software versus IRS-H

Test 2 determined whether the judgments made by the human aided by search software agree, and those made by IRS-H agree, thus testing the second hypothesis. Table 3 presents the results of the 5,200 judgement cases analysed testing sixty-five queries in eighty documents.

Table 3 Human judgements aided by search software versus IRS-H

Judgement test	Query structure	No of docs	No of cases	Control group	Kappa coefficient (*k*)	Strength of agreement	Significance (*p*)	Statistically significant
Human aided by search software * IRS-H	65 single queries each containing a single phrase-term	80	5200	IRS-H	1.000	Almost perfect	p < 0.001	Yes

Test 2 was used to test *$H2_0$* by looking at the strength of agreement between the judgements made by the participant in Experiment 3 and those judgements made by IRS-H in Experiment 2. The Kappa coefficient (k) and significance level (p) were calculated using SPSS and a six-division range to measure the strength of judgement agreements was used. The results revealed a Kappa coefficient of k = 1.000 with a statistical significance of p < 0.001 and a strength of agreement of 'Almost perfect'. The results from the test show that the agreement in judgements between IRS-H and the participant aided by search software was almost perfect. In all cases, agreements were made and the alternate hypothesis, $H2_1$ was therefore accepted that *the relevancy judgments made by a human aided by search software and an IRS using the hybrid indexing method agree.* Relevancy judgments made by a human aided by search software and an IRS using the hybrid indexing method disagree.

5.3 Strength of agreement scale

It must be noted that the six-division range for strength of agreement scale of Landis and Koch (1977:7) is misleading when k = 1.000, as the strength of agreement for this Kappa coefficient is described as 'almost perfect' when it should be described as 'perfect'. An updated version of their agreement scale is therefore proposed in Table 4.

Table 4 Updated strength of agreement scale (based on 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
1.00	Perfect

5.4 The vocabulary mismatch problem

The inner workings of IRS-H were known by the designer of the experiments and therefore the queries could be presented in the correct format to maximise system effectiveness. This worked as, in Test 2, the results showed that the judgements made by the participant, aided by search software, in all cases agreed with those judgements made by IRS-H. This suggests that there was no difference between the judgements of the two. When a match did occur, the vocabulary used in every query matched exactly that vocabulary that existed in a document. From a system perspective, IRS-H can therefore solve the problem of mismatching vocabulary between a query and a document. The first key IRS-H design feature that facilitates this is the use of the unique token 'identity' that maintains the word within the IRS-Hs token index, in the same order as they are read from the text. The second key design feature is that the phrase-terms are presented in the format IRS-H expects them to be in. This gives IRS-H the ability to match or not match queries to documents effectively. This effectiveness is supported by the participant, aided by search software, who agreed with all judgements made by IRS-H. However, the participants unaided by search software in Test 1 disagreed on many occasions; they considered that IRS-H made its judgements incorrectly compared with their own judgements. These judgments relate to the participants either stating the phrase-terms existed when they did not or stating the phrase-terms did not exist when in fact they did. The results suggest that, in this study, the participants made incorrect judgements when answering the questionnaire and that IRS-H, when presented with the correctly formatted queries, can make better judgements than the participants.

Returning to the busy general practitioner in the British Library: using an IRS with the hybrid indexing method on his document collection would show him how effective the method is. However, he would need to think carefully about what phrase-terms to choose before searching, and make sure those synonymic phrases are catered for to ensure relevant documents pertaining to his topic are retrieved and non-relevant documents are not.

6 Conclusion

Judgments made by humans and an IRS using the hybrid indexing method are not the same. In this study, there was a discrepancy between the outcomes as obtained by the participants and the IRS. The agreement or similarity of the findings between an IRS and humans is only moderate. When using an IRS on its own, the hybrid indexing method solves the problem of mismatching vocabulary between a query and a document. When participants use search software, the search results improve significantly. The agreement or similarity of the findings between humans using search software and IRS-H is almost perfect. It is concluded that the participant aided by search software or IRS-H improves the matching of vocabulary between a query and a document. However, IRS-H can perform the process rapidly on a large, closed document collection unassisted by the human. This combination will assist a researcher with a large collection of documents to search the closed collection accurately and quickly, producing quality results effectively.

This study has evidenced that an IRS, utilising the hybrid indexing method, can search for multi-word phrase-terms expressed within a query, and match these phrase-terms to those that exist in the text within a closed collection of documents. In its design, this method has the ability to extract whole paragraphs from text. This ability is extremely useful when searching for strings of text to quote references or when searching for products, names, diseases, codes, acronyms and quotations. It is important to note that the hybrid indexing method is not limited to the English language as it can be used in any language using the twenty-six-letter Latin-based alphabet.

References

Agnihotri, D., Verma, K. and 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.

Babbie, E. 2013. *The practice of social research.* 13th ed. Belmont, CA: Wadsworth.

Bauer, R.S., Brassil, D., Hogan, C., Taranto, G. and Brown, J.S. 2009. Impedance matching of humans <-> machines in high-Q information retrieval systems. *Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics.* 11-14 October 2009. San Antonio, TX, USA: IEEE SMC.

Bell, T.C., Moffat, A., Nevill-Manning, C.G., Witten, I.H. and Zobel, J. 1993. Data compression in full-text retrieval systems. *Journal of the American Society for Information Science*, 44(9): 508–531.

Binkley, D. and Lawrie, D. 2015. The impact of vocabulary normalisation. *Journal of Software: Evolution and Process,* 27(4): 255–273.

Binkley, D., Lawrie, D. and Uehlinger, C. 2012. Vocabulary normalization improves IR-based concept location. *Proceedings of the 28th IEEE International Conference on Software Maintenance.* 23-28 September 2012. Trento, Italy: ICSM. 588–591.

Burrell, G. and Morgan, G. 1979. *Sociological paradigms and organisational analysis.* London: Heinemann.

Bytheway, A.J. 2014. *Investing in information: the information management body of knowledge*. Switzerland: Springer.

Chaparro, O., Florez, J.M. and Marcus, A. 2016. On the vocabulary agreement in software issue descriptions. *Proceedings of the 2016 IEEE International Conference on Software Maintenance and Evolution.* 2-7 October 2016. Raleigh, NC, USA: ICSME. DOI:10.14738/tmlai.53.3106.

Clarke, C.L.A., Cormack, G.V. and Tudhope, E.A. 2000. Relevance ranking for one to three term queries. *Information Processing and Management*, 36(2): 291–311.

Cleverdon, C.W. 1967. The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6): 173-194. DOI: 10.1108/eb050097.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Croft, W.B. 2019. The importance of interaction for information retrieval. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 21-25 July 2019. Paris, France: SIGIR.

Croft, W.B., Metzler, D. and Strohman, T. 2015. *Search engines: information retrieval in practice.* Harlow: Pearson Education.

De Raadt, A., Warrens, M.J., Bosker, R.J. and Kiers, H.A.L. 2019. Kappa coefficients for missing data. *Educational and Psychological Measurement*, 79(3): 558–576.

Egozi, O., Markovitch, S. and Gabrilovich, E. 2000. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems*, 29(2): 1–34.

Fitzgerald, K.A., Fitzgerald, J.A. and Bytheway, A.J. 2017. Diabetes information retrieval research. *Journal of Health Informatics in Africa*, 4(2): 1–11.

Fleiss, J.L., Levin, B. and Paik, M.C. 2003. *Statistical methods for rates and proportions*. 3rd ed. Chichester: John Wiley & Sons.

Frej, J., Chevallet, J.P. and Schwab, D. 2018. Enhancing translation language models with word embedding for information retrieval. *arXiv,* 1801.03844.

Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11): 964–971.

Gale, E.A.M. 2001. The discovery of type 1 diabetes. *Diabetes*, 50(2): 217–226.

Gale, E.A.M. 2006. Declassifying diabetes. *Diabetologia*, 49(9): 1989–1995.

Gregor, S. and Hevner, A.R. 2013. Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2): 337–355.

Gross, P.L.K. and Gross, E.M. 1927. College libraries and chemical education. *Science*, 66(1713): 385–389.

Ha, L.Q., Sicilia-Garcia, E.I., Ming, J. and Smith, F.J. 2002. Extension of Zipf's law to words and phrases. *Proceedings of the 19th International Conference on Computational Linguistics.* 26-30 August 2002. Taipei, Taiwan: COLING. 1-6.

Harris, Z.S. 1954. Distributional structure. *Word*, 10(2-3): 146–162

He, B. and Ounis, I. 2009. Studying query expansion effectiveness. *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval.* 6-9 April 2009. Toulouse, France: ECIR

Hevner, A., Vom Brocke, J. and Maedche, A. 2019. Roles of digital innovation in design science research. *Business & Information Systems Engineering*, 61: 3–8.

Hevner, A.R., March, S.T., Park, J. and Ram, S. 2004. Design science in information systems research. *MIS Quarterly*, 28(1): 75–105.

IJzereef, L., Kamps, J. and De Rijke, M. 2005. Biomedical retrieval: how can a thesaurus help? In *On the move to meaningful internet systems 2005: CoopIS, DOA, and ODBASE.* Lecture Notes in Computer Science, 3761. R. Meersman and Z. Tari, Eds. Berlin, Heidelberg: Springer. 1432-1448.

Joyce, J. 1932. *Ulysses.* Reprint of 1932 edition. Ware: Wordsworth Editions Limited.

Jun, H.S. and Yoon, J.W. 2002. A new look at viruses in type 1 diabetes. *Diabetes Metabolism Research and Reviews*, 19(1): 8–31.

Katahira, M. 2009. A proposal for a new classification of Type 1 Diabetes Mellitus based on clinical and immunological evidence. *Recent Patents on Endocrine, Metabolic & Immune Drug Discovery,* 3(1): 54–59.

Kobayashi, V., Mol, S. and Kismihók, G. 2015. Discovering learning antecedents in learning analytics literature. (Unpublished).

Kohavi, R. and Provost, F. 1998. Glossary of terms. *Machine Learning*, 30(2/3): 271–274.

Koopman, B. and Zuccon, G. 2019. WSDM 2019 tutorial on health search: a full day from consumers to clinicians. *Proceedings of the 12th ACM International Conference on Web Search and Data Mining.* 11-15 February 2019. Melbourne, Australia: WSDM.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L. and Lawley, M. 2016. Information retrieval as semantic inference: a graph inference model applied to medical search. *Information Retrieval*, 19(1): 6–37.

Landis, J.R. and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174.

Langville, A.N. and Meyer, C.D. 2007. Information retrieval and Web search. In *Discrete mathematics and its applications: handbook of linear algebra.* K.H. Rosen and L. Hogben, Eds. Boca Raton, FL: Chapman & Hall/CRC. 63.1–63.16.

Liu, M., Fang, Y., Choulos, A.G., Park, D.H. and Hu, X. 2017. Product review summarisation through question retrieval and diversification. *Information Retrieval Journal*, 20(6): 575–605.

Manning, C., Nayak, P. and Raghavan, P. 2017. *Introduction to information retrieval – CS276 information retrieval and Web search – efficient scoring.* Stanford University. [Online]. https://web.stanford.edu/class/cs276/handouts/efficient_scoring_cs276_2013_6.pdf (11 October 2018).

Manning, C.D., Raghavan, P. and Schütze, H. 2008. *An introduction to information retrieval*. New York, NY: Cambridge University Press.

Markey, N. 2009. Tame the BeaST: the B to X of BibTEX. [Online]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.169.9276 (20 June 2019).

Min, J., Leveling, J., Zhou, D. and Jones, G.J.F. 2010. Document expansion for image retrieval. *Proceedings of the 9th RIAO Conference Adaptivity, Personalisation and Fusion of Heterogeneous Information.* 28-30 April 2010. Paris, France: RIAO. 65–71.

Mitra, B., Diaz, F. and Craswell, N. 2017. Learning to match using local and distributed representations of text for Web search. *Proceedings of the 26th International World Wide Web Conference.* 3-7 April 2017. Perth, Australia: IW3C2. 1291–1299.

Muller, H. and Holzinger, A. 2019. Kandinsky patterns. (Unpublished).

Nguyen, G.H., Tamine, L., Soulier, L. and Souf, N. 2018. A tri-partite neural document language model for semantic information retrieval. *Proceedings of the 15th Extended Semantic Web Conference.* 3-7 June 2018. Heraklion, Greece: ESWC. 445–461.

Onal, K.D., Zhang, Y., Altingovde, I.S., Rahman, M.M., Karagoz, P., Braylan, A., Dang, B., Chang, H. and Kim, H. 2018. Neural information retrieval at the end of the early years. *Information Retrieval*, 21(2-3): 111–182.

Panigrahi, D. and Gollapudi, S. 2013. Document selection for tiered indexing in commerce search. *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. 4-8 February 2013. Rome, Italy: WSDM. 73–82.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program*, 14(3): 130–137.

Rother, K.I. and Harlan, D.M. 2004. Challenges facing islet transplantation for the treatment of type 1 diabetes mellitus. *Journal of Clinical Investigation*, 114(7): 877–883.

Ruthven, I. and Lalmas, M. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2): 95–145.

Saunders, M., Lewis, P. and Thornhill, A. 2019. *Research methods for business students*. 8th ed. Harlow: Pearson Education.

Shekarpour, S., Marx, E., Auer, S. and Sheth, A. 2017. RQUERY: rewriting natural language queries on knowledge graphs to alleviate the vocabulary mismatch problem. *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. 4-9 February 2017. San Francisco, CA, USA: AAAI.

Sirres, R., Bissyandé, T.F., Kim, D., Lo, D., Klein, J., Kim, K. and Le Traon, Y. 2018. Augmenting and structuring user queries to support efficient free-form code search. *Empirical Software Engineering*, 23(5): 2622–2654. DOI: 10.1007/s10664-017-9544-y

Transier, F. and Sanders, P. 2008. Out of the box phrase indexing. *Proceedings of the 15th International Symposium on String Processing and Information Retrieval*. 10-12 November 2008. Melbourne, Australia: SPIRE. 200–211.

Tsikrika, T. and 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. and Croft, W.B. 1991. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3): 187–222.

Van Gysel, C., De Rijke, M. and Kanoulas, E. 2017. Neural vector spaces for unsupervised information retrieval. (Unpublished).

Van Rijsbergen, C.J. 1979. *Information retrieval*. 2nd ed. London: Butterworths.

Waitelonis, J. 2018. Linked data supported information retrieval. D.Eng thesis. Karlsruhe Institute of Technology.

Wang, Y., Huang, H. and Feng, C. 2017. Query expansion based on a feedback concept model for microblog retrieval. *Proceedings of the 26th International Conference on World Wide Web.* 3-7 April 2017. Perth, Australia: WWW. 559–568.

Williams, H.E., Zobel, J. and Bahle, D. 2004. Fast phrase querying with combined indexes. *ACM Transactions on Information Systems*, 22(4): 573–594.

Zhao, L. and Callan, J. 2010. Term necessity prediction. *Proceedings of the 19th ACM conference on information and knowledge management.* 26-30 October 2010. Toronto, ON, Canada: CIKM. 259-268.