Title A Heuristic Information Retrieval Study:

An investigation of methods for enhanced searching of distributed data objects exploiting bidirectional relevance feedback

Name Panagiotis Petratos

A HEURISTIC INFORMATION RETRIEVAL STUDY:

An investigation of methods for enhanced searching of distributed data objects exploiting bidirectional relevance feedback

Panagiotis Petratos

A thesis submitted for the degree of Doctor of Philosophy
of the University of Luton

The University of Luton,
Park Square,
Luton,
Bedfordshire,
LU1 3JU.

February 2004

A HEURISTIC INFORMATION RETRIEVAL STUDY:

An investigation of methods for enhanced searching of distributed data objects exploiting bidirectional relevance feedback

Panagiotis Petratos

ABSTRACT

The primary aim of this research is to investigate methods of improving the effectiveness of current information retrieval systems. This aim can be achieved by accomplishing numerous supporting objectives.

A foundational objective is to introduce a novel bidirectional, symmetrical fuzzy logic theory which may prove valuable to information retrieval, including internet searches of distributed data objects. A further objective is to design, implement and apply the novel theory to an experimental information retrieval system called *ANACALYPSE*, which automatically computes the relevance of a large number of unseen documents from expert relevance feedback on a small number of documents read.

A further objective is to define a methodology used in this work as an experimental information retrieval framework consisting of multiple tables including various formulae which allow a plethora of syntheses of similarity functions, term weights, relative term frequencies, document weights, bidirectional relevance feedback and history adjusted term weights.

The evaluation of bidirectional relevance feedback reveals a better correspondence between system ranking of documents and users' preferences than feedback free system ranking. The assessment of similarity functions reveals that the Cosine and Jaccard functions perform significantly better than the DotProduct and Overlap functions. The evaluation of history tracking of the documents visited from a root page reveals better system ranking of documents than tracking free information retrieval. The assessment of stemming reveals that system information retrieval performance remains unaffected, while stop word removal does not appear to be beneficial and can sometimes be harmful. The overall evaluation of the experimental information retrieval system in comparison to a leading edge commercial information retrieval system and also in comparison to the expert's golden standard of judged relevance according to established statistical correlation methods reveal enhanced system information retrieval effectiveness.

ACKNOWLEDGMENTS

The author is indebted and wishes to express his thanks for the extensive and intensive support and encouragement of all the friends and family during the years of study and research in the development of the work for this thesis.

Many people have helped in a variety of ways in the work leading up to this thesis. They include:

George Petratos

Ero Petratou

Dimitris Petratos

Richard Forsyth

Gordon Rugg

Alfred Vella

Angus Duncan

Martha Pennington

Li Chen

The author is also indebted and wishes to express his appreciation to all the anonymous referees acknowledging their valuable suggestions for improving all the associated published work and consequently this thesis.

LIST OF PUBLICATIONS

Petratos P., Chen L. and Wang P. "Bi-directional Fuzzy Logic Applied to Information Retrieval and Learning" in press Information Science Journal, 2004.

Petratos P. "A polythematic real-time synergistic hybrid data telecommunication system for scientific research with bidirectional fuzzy feedback peer review by expert referees" Data Science Journal, Vol. 2(4) pp: 47-58, 2003.

Petratos P., Chen L., Wang P., Forsyth R. "A Bi-directional Fuzzy Logic Theory: The Generalized Knuth's Triadic Logic for Information Retrieval". In Proceedings of the IEEE Systems Man and Cybernetics Conference, October 2002.

Petratos P., Chen L., "A note on bidirectional fuzzy logic". In Proceedings of the North American Fuzzy Information Processing Society Conference, June 2002.

Petratos P. and Vella A. "Heuristic algorithms for internet reduction" Synopsis based upon presentation given at the Operational Research International Conference at the University of Wales 12th September 2000.

Petratos P. and Vella A. "Electronic commerce via the internet, a view from the other side of the pond" Synopsis based upon presentation given at the Operational Research International Conference at the University of Lancaster 8th September 1998.

LIST OF CONTENTS

LIST OF CHAPTERS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 – Information retrieval systems: introduction, objectives, background and applications

1.1 Introduction and objectives

This chapter is the prologue of this thesis and presents an introduction to information retrieval systems. The origins, scope, basic concepts, principles, techniques and applications of information retrieval systems are discussed.

Furthermore, an introduction to artificial intelligence is presented as the syntheses of information retrieval methodologies and artificial intelligence techniques are integral elements of this work. The primary aim of this work is to investigate methods of improving the effectiveness of current information retrieval systems.

This aim can be achieved by:

- Introducing a novel bidirectional, symmetrical fuzzy logic theory which may prove valuable to information retrieval, including internet searches of distributed data objects.
- Applying bidirectional fuzzy logic to classic information retrieval theory.
- Defining an experimental information retrieval methodology framework within which a variety of similarity functions, term weights, relative term frequencies, document weights, bidirectional relevance feedback and approaches to history tracking may be tested.
- Designing and implementing an academic experimental information retrieval system called *ANACALYPSE*, which uses bidirectional fuzzy logic to compute automatically the relevance of a large number of unseen documents from expert

relevance feedback on a small number of documents read and forms an essential component of the above mentioned experimental framework.

- Evaluating the impact of bidirectional relevance feedback, stemming, stop-word removal and history on the effectiveness of the experimental information retrieval system.
- Evaluating the effectiveness of the academic experimental information retrieval system in comparison to a commercial information retrieval system and comparing both to the expert's gold standard of judged relevance according to established statistical correlation methods.

This chapter provides an overall view of the whole work of this thesis. Section 1.7 of this chapter presents the structure of the work of this thesis as analyzed in the series of subsequent chapters.

1.2 Information retrieval systems background

Information retrieval is a science which during the twenty first century has attracted increased interest principally due to the necessity of discovering relevant information in this modern age of information overabundance. It is true now more than ever before that this information overabundance is leading to a congestive information overload.

The ensuing number of publicly available documents and data objects of all types of stored information is rising rapidly. However, even though the quantity of information is constantly expanding there is no clear evidence that there is an analogous correlation with the *quality* of this overabundance of information.

Furthermore there is clear evidence reported recently that there is actually very little *new* information in all this overabundance of information produced annually (Lyman and Varian, 2003). In fact there are numerous scholars and scientists, including the author, who believe that the quality of scientific publications unfortunately does not increase with the ever increasing quantity (Baeza-Yates and Ribeiro-Neto, 1999).

Heretofore continuous technological developments in digital microelectronics have led to the sustained growth of computing power necessary for real time pattern recognition systems (Adjei and Mrozek, 1999; Adjei and Vella, 2000). Furthermore continuous technological developments in computing have led to the sustained growth of data storage capacity and network communication capacity. These advances subsequently led to the rapid world wide growth of digital electronic information and consequently to the widespread usage of information retrieval systems.

Information retrieval systems are the protagonists of information science and hold a critical role in library science and in general in computer science. The phrase information retrieval was first introduced to the academic literature in a research paper by Mooers (Mooers, 1952) and approximately a decade later was further supported and disseminated through the work of Fairthorne (Fairthorne, 1961).

Furthermore, the value and importance of information retrieval systems are well recognized in all fields of science with information seeking requirements and communication needs (Bessis, 2003a; 2003b). A large number of paradigms of information retrieval systems for various applications form a global sphere of

influence recognized in all scientific disciplines that require access to distributed textual or multimedia information in disseminated databases.

The exact functionality of a database system depends critically upon the characteristics of the information that it stores and searches. Therefore this is the primary aetiology for the ensuing development of various types of database systems. Information retrieval systems utilize a database which is essentially a collection of documents.

A document is a sequence of terms, expressing ideas about some topic in a natural language. The unit element of a document is a term, or a word, or potentially a root of a word. The semantic unit of a document is a sentence, or a phrase and a query is a set of terms representing a request for documents pertaining to some topic. Information retrieval systems perform automated retrieval of documents with information content relevant to a user's information search request.

The primary difference between information retrieval systems and database management systems is that whilst the former process unstructured free text data the latter process structured data. Furthermore, even with modern information retrieval systems, if large numbers of queries are issued on their databases, they operate at an average accuracy which does not exceed 50% (Eastman and Jansen, 2003).

Although this is a considerable improvement over earlier information retrieval systems which did not exceed 30% average accuracy (Cleverdon, 1972) nonetheless modern information retrieval systems' average accuracy still remains consistently

lower when compared to database management systems which operate about at 100% accuracy as shown in Table 1.1.

Naturally, this discrepancy in performance is due to the fact that information retrieval systems search their database and respond to users based on the relevance of the content of the documents to the queries. Hence in information retrieval systems search relevance is characterized by degrees of precision, whereas in database management systems search relevance is characterized by exactness.

For instance, the query: find the customer with the highest number of orders, has an exact answer, whereas the query: find the document most relevant to the life cycle of the hippopotamus has no exact answer and is rather subjective.

Table 1.1. Information retrieval and database management systems characteristics.

Characteristics	Information Retrieval Systems	Database Management Systems
Data	Unstructured Free Text	Structured Records
Data Volumes	Web Capacity	Few Terabytes
Architecture	Distributed widely with myriads of sporadic clusters	Centralised Control
Query Language	Natural Language	SQL
Operators	Relevance, Near, Similar, etc.	Select, Join, Union, etc.
Techniques	Heuristics	Search Algorithms
Search Relevance	Semantic Approximation	Exactness
Accuracy Average	50%	100%
Validation	Subjective	Objective

Hence information retrieval systems store and search information in documents (Smith and Devine, 1985) authored in natural language whilst database management systems store and search in records with strictly numeric or restricted language information.

However, contemporary documents are not mere text anymore. With the development and progression of the internet documents have evolved into polymorphous data objects containing information in a variety of data types and formats including text, graphics, images, sound and video. In the work of this thesis, attention is drawn to this type of document in a distributed information retrieval environment.

1.3 Information statistics and retrieval problems

The internet is the newest of all information communication channels including all types of broadcast and telephony information channels. Internet users increasingly find it indispensable while its growth shows that it is the fastest increasing new information channel of all time.

Furthermore there are four types of stored information, paper, film, magnetic and optical. It is noteworthy that the internet can combine all types of stored information unlike the other information channels. The internet includes two variants of the world wide web, the so called surface web which includes the open static web pages and the deep web which includes the database driven dynamically created web pages as shown in Table 1.2.

Table 1.2. Information size of the web in terabytes (Lyman and Varian, 2003).

Information Channel	Size in Terabytes
Surface Web	167
Deep Web	91,850
Total	92,017

Another approach to estimate the size of the information in the web relies on the number of hosts connected to the internet. Although this approach does not provide a definitive answer for information size, researchers utilize statistical techniques whereby a statistical sampling study is employed in order to estimate the size of an average web page and the contents of an average web site belonging to each host (Lyman and Varian, 2003). There are numerous methods for counting the number of hosts which differ in frequency and strategy according to the research company conducting the study, as listed in Table 1.3.

Table 1.3. Information size of web estimation methods (Lyman and Varian, 2003).

Company	Domains	Method	Frequency
www.whois.net	31,987,198	Domain name registration	Continuous
www.netcraft.com	42,800,000	HTTP requests, responses	Monthly
wcp.oclc.org	9,040,000	IP addresses (1 has many Domains)	Annually

The difference of the estimated number of domains between www.whois.net and www.netcraft.com is due to the fact that many web servers correspond to virtual names hence the first company lists them as one whereas the second lists them as many.

The difference between wcp.oclc.org and the other two companies regarding the estimated number of hosts is due to the fact that the wcp.oclc.org company only counts IP addresses each of which are capable of containing a large number of domain names.

There are interesting search statistics which show the total time spent by all visitors searching at each engine and the number of searches issued per day for each search engine as shown in Table 1.4. Furthermore, the search provider makes available its own database to a search engine.

The numbers of searches issued to a search engine and the amount it is used by searchers as well as other search engines are measures which imply quality. The top search engine from a searchers' utilisation point of view is Google.

Furthermore, it is noteworthy that in the top five search engines from a searchers' utilisation point of view, the four search engines make use of Google as their search provider.

Therefore, in the work of this thesis attention is drawn to the development and evaluation of an academic experimental information retrieval system in a comparative study in contrast to Google.

Table 1.4. Information search statistics on the web (Lyman and Varian, 2003).

Search engine	Search hours/month in millions	Searches/day in millions	Search provider
Google	18.7	112	Google
AOL	15.5	93	Google
Yahoo	7.1	42	Google
MSN	5.4	32	Overture
Ask Jeeves	2.3	14	Google
InfoSpace	1.1	7	Google
Altavista	0.8	5	Overture
Overture	0.8	5	Overture
Netscape	0.7	4	Google
Earthlink	0.4	3	Overture
Looksmart	0.2	1	Looksmart
Lycos	0.2	1	Overture
Total	53.2	319	

Table 1.4 shows in descending order the search engines according to the total number of search hours which users spend per day and per month and for each search engine the corresponding search provider, Google in most cases at the top of the table.

Moreover, the total number of internet users is currently approximately 600 million. The users are evenly distributed around the globe. Approximately 30% are in North America which refers to USA and Canada, about 30% are in Europe and about 30% are in Asia Pacific including Australia and New Zealand. Latin America currently has

6% the largest share of the remaining users' distributions (Lyman and Varian, 2003). These statistics show that India and China are still in their infancy when it comes to internet access.

In the not so distant past readers relied on human experts familiar with the particular information environment, namely the librarians, in order to discover information relevant to their needs in a particular library. On numerous occasions an expert librarian would be able to act in response to the reader's information need with an unambiguous although unfortunately adverse reply to the reader's query.

Recently, with the enormous increase of all the books, journals and periodicals the reliability of the human expert method has diminished rendering almost impossible the existence of experts who can possess all the knowledge regarding everything published.

Thus, readers increasingly rely on machines to perform automatic information retrieval. However, this method brings to light other new challenges for the readers, such as appropriately expressed query formulation and questionable system effectiveness. Both factors culminate in very large number of results to read and evaluate of which only a few might be relevant for each individual reader. It is apparent that more efficient tools are required for modern information retrieval.

The work presented in this thesis is concerned with the investigation of methods for improving the effectiveness of current information retrieval systems. This investigation is carried out within an experimental methodology supported through

the development and implementation of a heuristic information retrieval system in order to alleviate a number of modern information retrieval problems.

A few of the current information retrieval problems include the following. The primary issue is the vast amounts of data ensuing from myriad distributed documents in the ever expanding internet environment which increases information retrieval complexity. Furthermore another issue is the ephemeral nature of the environment with individual documents and even entire domains disappearing and often reappearing at different locations with different names.

Also an additional issue is that new documents and even entire new domains are emerging constantly. Moreover changes are taking place at short time intervals often without any warning. In addition current information retrieval systems exhibit a low coverage of the available data set and high levels of computational complexity during the process of finding, retrieving, indexing and updating distributed documents in their database.

Furthermore current information retrieval systems exhibit weakness in fulfilling a user information search request with vast numbers of retrieved results which are impossible for the user to read in their entirety. Also current information retrieval systems tend to retrieve gigantic answer sets which are highly ambiguous semantically and often inadequate to satisfy an expressed user information need.

The work of this thesis presents a number of new techniques developed to address some of the aforementioned problems. These include the introduction of a bidirectional fuzzy logic theory in synthesis with information retrieval methodologies.

1.4 Basic system model

An information retrieval system is designed to seek, find and retrieve relevant information in response to a human information search request. In that sense information retrieval systems are essential tools of scientific research as more and more knowledge regarding all scientific disciplines is transferred from print to digital form.

The basic idea is to model information retrieval systems such that they are in a position to match the relevance of a search request to documents in a database on a functional basis as closely as possible to a human reader. This of course presents two different but much related problems.

The first problem is the characterization of the searchers' information request at the input interface of the information retrieval system. For instance, consider an information search request such as find all the recent Journal articles which do not exceed the age of three years old within the information retrieval research area written in the English language by authors of Hellenic descent.

Unfortunately this information search request which is expressed in natural language cannot be used directly as input via the current interfaces of information retrieval systems. Hence, the searcher must first provide an acceptable translation which is

understandable by the information retrieval system as a replacement for the information search request expressed in natural language.

This translation is also known as the query and consists of a set of keywords which form a synopsis of the full description of the information search request expressed in natural language. Therefore, if the input to an information retrieval system is a searcher defined query then the output is a results list of retrieved information ranked in order of importance which might be useful and relevant to the searchers' request.

The second problem lies in the heart of the information retrieval system and is concerned with the machine understanding of the texts in the database in order to determine which texts are the most relevant to the searchers' information request.

For instance, lion kills man and man kills lion, are two phrases which consist of the exact same terms however they carry completely different, opposite meanings. These antitheses of the meanings are indicated simply by the different order in which the terms appear in each phrase.

Although this is a simple example it clearly shows a few of the difficulties and challenges in machine text understanding and furthermore in machine information retrieval. Hence the information retrieval task of determining which texts in a database contain the same keywords as in the searcher's query most frequently is not enough to satisfy the searcher's information request.

In order to satisfy the searcher's information request the contents of the texts in a database must be understood by the information retrieval machine by extracting syntactic and semantic information from the texts and matching it to the searcher's information request so that the texts can be ranked according to relevance.

Naturally the difficulty and the challenge are not only to know how to extract semantic information from the texts but also how to use it to judge relevance. Therefore the idea of relevance lies in the heart of information retrieval.

Although in the area of machine text recognition significant progress has been made to the point where optical character recognition systems operate at a high enough accuracy to emulate human vision for reading, still human-like text understanding remains elusive for machines. Therefore, since human cognition is a complex mechanism, machine information retrieval is considered to be an approximation of the human ability to read and understand information.

Replicating human cognition of relevance requires immense research efforts toward technological developments and prolonged research in various scientific areas including biology, medical science, neuroscience, cognitive science and computer science. Hence, heretofore an information retrieval system can be modelled as a series of computational stages and processes with more modest aspirations and rather pragmatic aims.

A basic information retrieval system is illustrated in the following figure. This basic model comprises of a number of stages: the searcher's query formation and issue

stage, the system's database search and query matching stage, the system's ranking of the results stage and the searcher's browsing of the results stage. From the point of view of the searcher out of all these stages the most costly stage in terms of time and effort is the activity of browsing.

If luck is on their side, searchers will stumble upon the missing piece of information immediately or very quickly. However, most frequently the searcher has an information request which is either poorly defined as a query or is inherently broad as a subject.

For instance, consider a case in which the searcher might be interested in ancient Hellenic artefacts in general without being an expert on Hellenic archaeology. Hence, in such a case an interactive interface might be used by the searcher to simply look around in the database for documents related to archaic Hellenic artefacts.

The searcher might find interesting documents about the Athenian, Peloponnesian, Cycladic and Minoan civilisations. Furthermore, while reading about the Minoan civilisation the searcher might follow interesting hyperlinks to a map including directions to the palace of Knossos and from there the searcher might follow interesting hyperlinks to tourist information about Crete and from there to tourist information about Hellas.

Such a case is characterized by the activity of browsing whereby the searcher indulges into a journey of visiting and reading interesting subsequent hyperlinked information. This is a case of information retrieval whereby the searcher's principal objectives are not clearly defined from the start and the purpose of search might change during the

session and according to the searcher's interaction with the information retrieval system.

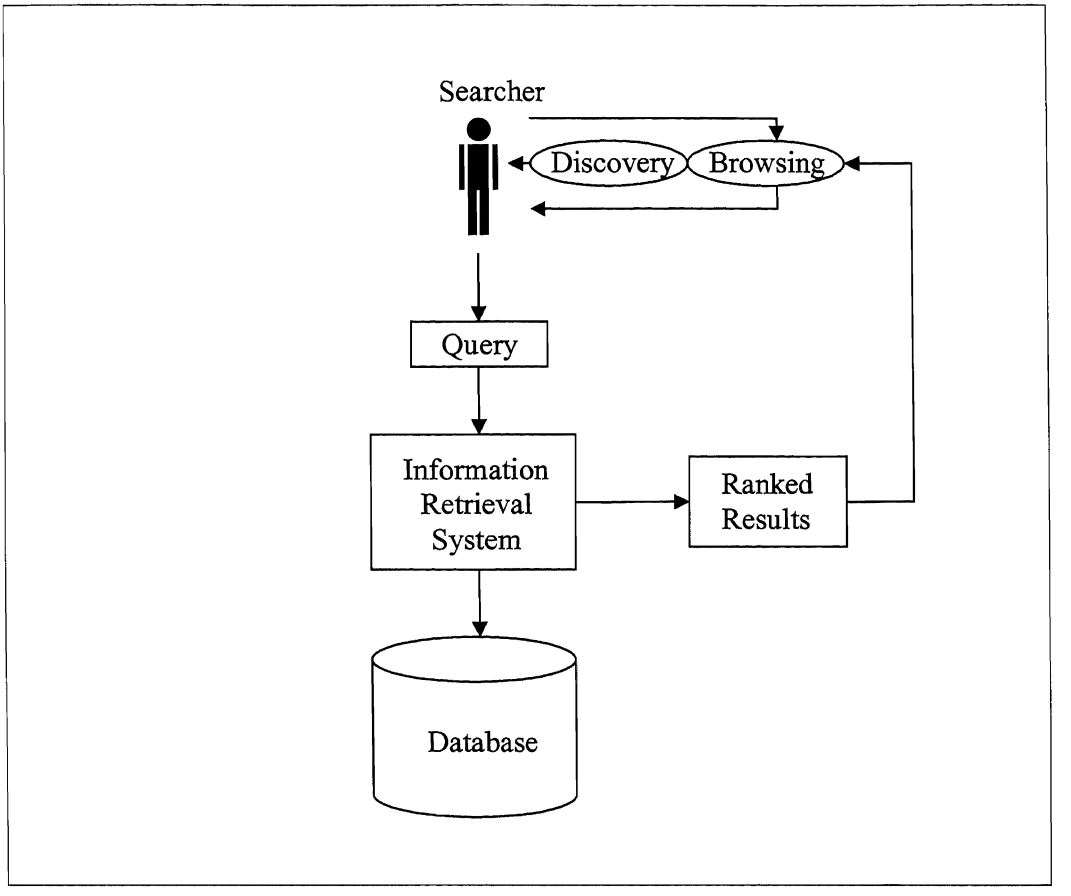

Figure 1.1. Basic model of an information retrieval system.

All the fundamental components displayed in Figure 1.1 are the most essential elements of an information retrieval system. Naturally their individual operating conditions or the variations of the design of any of these elements have a considerable impact on the design of the information retrieval system on the whole.

For instance if instead of having a local centralised database such as in a library, the databases are distributed then careful consideration should be given to the design of

the remote connection, access, retrieval and processing algorithms as well as to the underlying communications network design.

In addition, if a large number of searchers are anticipated to utilize the information retrieval system then careful consideration should be given to the concurrency design of the system algorithms. Furthermore, automated categorisation of texts into predefined categories is a whole research area of interest which is mainly concerned with either the classification or elimination of a document as belonging to a particular group of documents.

Also, text categorisation according to author is another research area of great interest and lends itself harmonically to identifying the authors of texts of previously unknown ownership which is also known as authorship attribution (Forsyth, 1995). Hence from the categorisation point of view information retrieval is a classification problem elucidated at a very fine level of semantic class which is based upon the expressed information request of the searcher.

Information retrieval systems have been used for a variety of applications. Examples of such applications include medical, scientific and industrial uses. Information retrieval systems can be implemented in both computing hardware and software.

Computer hardware accelerated processing can generate considerable advantages in terms of the speed and accuracy of the system however this strategy often narrows and limits the scope of the system while incurring a considerable ensuing financial cost.

The work presented in this thesis focuses on the development of a system using software approaches which provide greater flexibility for further developments at a lower cost.

1.5 Applications of information retrieval

Heretofore researchers have pursued the development and application of many types of information retrieval systems for a wide range of areas. Examples of such systems that have been applied to scientific, industrial, medical, legal and social uses are a few of the paradigms of information retrieval systems applications. The genesis of information retrieval systems took place in the libraries where they were used for searching and retrieving records from bibliographic databases (Frakes and Baeza-Yates, 1992).

Initially, due to computer hardware limitations early information retrieval systems instead of accessing entire documents they only provided access to references such as the ones found in bibliographies at the end of documents (Sparck Jones and Willett, 1997). Also, for the same reasons early information retrieval systems performed indexing and searching on document surrogates such as titles and abstracts instead of the documents themselves.

Naturally technology has progressed a long way since then and currently not only full text search and retrieval is available on a variety of topics and document types which can be stored in local or distributed databases but in addition contemporary information retrieval is performed with a greater ease of use as searchers interact with

the systems through more natural user interfaces such as voice (Sparck Jones, Jones, Foote and Young 1997; Smith and Linggard, 1982).

1.6 Introduction to artificial intelligence and bidirectional fuzzy logic

Artificial intelligence as a phrase was introduced in 1956 when researchers John McCarthy, Martin Minsky and Claude Shannon organised a summer workshop with the same title at Dartmouth College which at the time was home to John McCarthy.

This workshop was aimed at researchers interested in cybernetics, machine intelligence, artificial neural nets and automata theory portraying a new science called artificial intelligence.

Almost a decade later Lotfi Zadeh, a professor from the University of California at Berkeley published his well known paper entitled *fuzzy sets*, introducing the theory of fuzzy logic (Zadeh, 1965).

Lotfi Zadeh in his seminal research paper *fuzzy sets* proposed that a fuzzy set A is characterized by a membership function mapping the elements of a domain, space, or universe of discourse X to the unit interval $[0.0, 1.0]$ expressing the set membership function as $\mu_A : X \rightarrow [0.0, 1.0]$.

To introduce bidirectional fuzzy logic consider that if Zadeh's definition is modified to include the additional symmetrical negative space then a fuzzy set A is characterized by a membership function mapping the elements of a domain, space, or universe of discourse X to both positive and negative unit intervals

$[-1.0, 0.0] \cup [0.0, 1.0]$ expressing the set membership function in the following manner $\mu_A : X \rightarrow [-1.0, 0.0] \cup [0.0, 1.0]$.

In which case the value of $\mu_A(x)$ for the fuzzy set A is the membership value or the grade of membership of $x \in X$. Ergo, the element membership value illustrates the degree of x belonging to the fuzzy set A.

The 1960s was a time of euphoria for artificial intelligence as it attracted the interests of great scientists and researchers who introduced fundamental new ideas in knowledge representation, learning algorithms, fuzzy logic, neural computing and computing with words (Negnevitsky, 2001).

Unfortunately due to computer hardware limitations at the time a lot of these new ideas either could not be implemented or they could not be developed to their full potential. Artificial intelligence is a very ambitious idea because it aims to create machines which exhibit a very human characteristic, intelligence.

Hence, artificial intelligence inspires researchers and fills them with euphoria of high-flying expectations. However such a grand ambition also incubates the danger of failure of not meeting such exceedingly grand expectations.

Artificial intelligence's high-flying expectations imperilled falling from grace metaphorically just like Icarus fell from his high flight to his demise in the well known allegory of Hellenic mythology.

In the early 1970s the euphoria about artificial intelligence was replaced by disillusionment and frustration as the achievements were limited to machines playing games with few practical applications as no artificial intelligence system was capable to successfully solve any real world problems (Russell and Norvig, 2002).

Interestingly enough as the capabilities of computers grew stronger artificial intelligence researchers focused on a few of the more pragmatic aims which were achieved initially with expert systems, then with neural networks, evolutionary computing, fuzzy logic and computing with words (Negnevitsky, 2001).

This paradigm shift reinvigorated the discipline and provided the appropriate environment for further increased research leading to new discoveries and novel, pragmatically valuable real world innovations.

1.7 Organisation of the thesis

The organisation of this thesis is as follows. Chapter two reviews the literature that covers work conducted on the topic of information retrieval and describes details reported in the literature regarding various information retrieval methods and developments in fuzzy information retrieval.

Chapter three discusses bidirectional fuzzy logic, its applications to information retrieval and a comparison of standard fuzzy logic and bidirectional fuzzy logic with representative paradigms.

Chapter four discusses the research methodology developed for the experimental work of this thesis. This research methodology relies on a framework comprising various functions and components which are designed according to the optimum experimental results observed in the literature while introducing novel functions.

Furthermore the research methodology is designed with the scope of investigating which of the novel functions introduced are the best for effective information retrieval results.

Chapter five discusses a comparative study of the work of this thesis presenting some of the experiments conducted in comparison to Google, the commercial information retrieval system. The concepts, theories, rationale and contributions of the work of this thesis are discussed.

Chapter six discusses the architecture of the academic experimental information retrieval system and outlines detailed descriptions of the system components. Furthermore the design philosophy of the new information retrieval system and how it relates to the novel theory introduced with this thesis is discussed.

This work expands upon the experimentation of chapter five by including the history feature which takes into account the free searcher's web roaming in order to investigate whether it makes a difference on the effectiveness of the information retrieval process.

Chapter seven presents conclusions derived from the work of this thesis. Suggestions for future work are discussed. Furthermore this chapter raises important points observed from the outcome of experiments for effective information retrieval.

CHAPTER 2 – Review of previous work

2.1 Introduction

This chapter reviews the related previous work regarding information retrieval systems including pertinent methodologies involving artificial intelligence techniques. Attention is drawn to fuzzy information retrieval systems, as an important objective of the present work is to implement and develop an academic experimental information retrieval system with bidirectional fuzzy relevance feedback. It is hoped that a few of the gaps in current research will become clearer through this discussion with respect to the work of this thesis.

2.2 Historical survey of information retrieval

For more than 4000 years there has been a need for organization of information in order to facilitate manageable retrieval. This fact is true not only for the modern information seekers but it was also true for the archaic readers and authors even though the amount of published information was much less than at present and the archaic information retrieval tools were much simpler compared to modern tools.

The earliest and most important information retrieval tools invented were the table of contents and the index both of which equally enjoy a great popularity with the public even to this day. The table of contents holds a synopsis of the semantic pith of the book and the locations of the principal ideas within it.

The index holds an alphabetical list of key words and phrases with their corresponding page numbers which are used as reference points to find where

information about these key words appears in the text. Naturally as humans are prolific authors and fanatical readers the amount of published information eventually surpassed the number of a few books which created a need for specialised data structures to ensure manageable organisation and retrieval of all the published information.

Hence, the good old index was expanded in order to include from a collection of books, key words and their associated pointers -books and page numbers- to their related information in the texts. The Sumerian literary catalogue, of circa 2000 B.C., is not unlikely to be the first list of books ever written (Frakes and Baeza-Yates, 1992).

The range of key words in a modern index can be quite broad, including authors and key concepts, typically appearing at the last few pages of a book immediately following bibliography where relevant cited publications are listed such as books, journals, conference papers and research theses.

Modern book indexes evolved from an archaic form which started appearing in the early sixteenth century and eventually developed to a homologous form not unlike the modern index during the late eighteenth century. Libraries utilize a different type of index which appears in the form of card collections or library catalogues.

For instance, in 1604 the Bodleian library catalogue at Oxford University was not unlikely the largest of its kind covering of the order of 10,000 books written on a variety of subjects by well known authors such as Hippocrates, Euclid, Plato and

Aristotle (Rogers, 1995). However, index creation at the time was neither a quick nor an easy task to accomplish as all the required work was conducted manually by humans.

A few centuries later during the dawn of the twentieth century the first edition of the concordance of the Hellenic New Testament was published by William Moulton and Alfred Geden in 1897. Since then numerous revisions have followed and as a family labour of love they were undertaken by three generations of the Moulton descendants (Witten, Moffat and Bell, 1999).

Approximately a decade later the concordance of the complete poetic anthology by William Wordsworth was published in 1911 by Professor Lane Cooper. However this was neither an individual accomplishment nor a painless achievement. This tome consisted of 1,136 pages and included 211,000 not so trivial key words. This concordance was manually completed by a highly organized team of sixty-seven people three of whom had died by the time the concordance was finally published (Witten, Moffat and Bell, 1999).

It is quite clear that concordance development and index formation until the early twentieth century has been a rather laborious, onerous, manual and monotonously prolonged procedure. During that period information retrieval was also a manual process performed by reading the document references written typically in a collection of cards or library catalogues and discovering the relevant documents by trial and error.

However, in the early nineteen fifties a metamorphosis for information retrieval was quietly taking place as computers were appearing with increasing capabilities of electronically recording and storing document references in databases and gradually eliminating the need for manually created cards or library catalogues.

As computers were assuming an increasingly important role for the academic world the computer science field of information retrieval progressively gained broad acceptance among scholars and scientists which has never before been more evident than in the late twentieth and early twenty first century, the age of information overabundance.

However, in its early stages electronic information retrieval was concerned with automatic data processing which distinctively signifies the actions of recording and storing surrogate information about the book such as authors names, title, abstract, number of pages, and a reference of where the book is physically located in the library.

Early information retrieval systems instead of preserving the whole book itself electronically, only allowed a glimpse to the reader who could just only access the book of interest by obtaining an actual library hard copy. Therefore, during that period information seekers were restricted to retrieval of surrogate data rather than being allowed to have the benefit of full text retrieval of the actual information of interest.

During that epoch it was not pragmatically viable to preserve electronically all the information included in a document, first and foremost due to computing hardware resource limitations and second due to economic restrictions. Furthermore, numerous scholars at the time indicated the void of the computerised manuscript information absence and emphasized the critical importance of filling this chasm with the ability to electronically store the whole document (Brooks and Iverson, 1963).

It was not long before this next critically important milestone became a reality and information retrieval systems had the capability of electronically storing paradigmatic data structures combining key words with a text compression scheme symbolizing the full contents of a document in order to have the ability to comprehensively perform broad ranging information retrieval.

Ensuing with these developments a formal information retrieval definition elucidated that an information retrieval system provides information about a service or a situation whenever and wherever this information may be required (Martin, 1967).

Therefore, from the a priori definition it may be further inferred that the most important aim of an information retrieval system is to make available information to the seeker which is precisely relevant with respect to an explicitly expressed user information need.

During the nineteen seventies the capabilities of information retrieval systems were increasingly expanded in order to store and retrieve steadily growing data structures of the comprehensive document information. However despite technological progress

during that period the majority of information retrieval systems still did not make available any detailed information beyond mere document references (Lancaster, 1978).

Approximately a decade later, during the nineteen eighties the next generation of information retrieval systems were largely investigated via the engagement of various natural language algorithms which treated document text as meaningful succession of concepts rather than regarding text simplistically just as character streams.

Natural language research also had a great positive impact on the further development of speech recognition (Smith and Clotworthy, 1988; McMahon and Smith, 1998). In tandem these two research areas have the potential to provide very powerful human computer interfaces for information retrieval systems (Smith and Linggard, 1982).

In the next decade, during the nineteen nineties the capabilities of computers increasingly expanded to make available to users not only full text analysis but also acoustical and optical processing of information of multiple types due mainly to previously prohibitive but now countermanded computer hardware limitations and rescinded economic restrictions.

Books, journals and conference papers are not anymore exclusively under the physical ownership of libraries but can also be found in electronic form under virtual libraries on the internet such as the *IEEE* or the *ACM* digital libraries and as the amount of accessible digital information has increased so have the expectations of readers for information retrieval systems.

Hence, the first priority of modern information retrieval systems is not anymore a mere reference to the physical location of the document of interest but instead the primary concern is relevant, precise information satisfying a user information need (Van Rijsbergen and Lalmas, 1996).

Thus, modern information retrieval is not just about searching for answers in electronic text but also about being in quest for knowledge in electronic information such as acoustical, optical, cinematographic, computer generated images, graphics and their optical animations.

Another very interesting recent approach to the information retrieval problem is intelligent software systems which have been pioneered in the field of artificial intelligence and are commonly referred to as intelligent agents (Estall and Smith, 1984; Busetta, Serani, Singh and Zini, 2001).

Intelligent agents are autonomous software systems a few of which are capable of natural phonetic or hand written human computer interaction which specialize in specific tasks, are able to determine the optimal course of action and then execute complicated procedures to achieve the desired result.

Intelligent agents are also exceptionally adaptable to distributed environments principally due to their increased autonomy. Intelligent agents with natural interaction interfaces and information retrieval capabilities possess a remarkable potential to

interact, reason and easily discover desired information even in a distributed collection of documents.

Distributed intelligent agents have often been engaged by researchers and commercial search engines for internet information retrieval with very encouraging results. Critics argue that intelligent agents promise much but question whether reality does match promise? (Wagner and Turban, 2002).

Naturally scientific progress is made not in giant leaps but in small incremental steps. Intelligent agents possess a potential for further artificial intelligence research and hold promise especially for the area of human computer interaction in synthesis with the field of information retrieval (Smith and Linggard, 1982; Russell and Norvig, 2002).

2.3 Historical survey of fuzzy logic

A priori development and a posteriori progress of logic was only possible principally due to the syllogisms of Aristotle and the Hellene philosophers who preceded him. Also, evolution of geometry and mathematics was only possible principally due to the contributions of Hellene mathematicians including Pythagoras, Thales and Euclid.

Even though modern logic is commonly credited to Aristotle who laid the foundations with his syllogistic logic there is some evidence that logic had been developed for quite some time before Aristotle by various preceding Hellene philosophers.

Parmenides was the earliest of Aristotle's predecessors who first around 480 B.C. proposed the law of the excluded middle which states that every proposition must be either true or false and based on that philosophy Parmenides established his own school of logic in Elea which was also known as the Eleatic school of philosophy (Devlin, 1998).

However, the teachings of Parmenides were not well received by all philosophers. Heraclitus notably theorized that in certain situations events could be at the same time true and the antithetical not true in this cosmos which is characterized by constant change.

Another well known Hellene philosopher was Plato who founded in Athens around 387 B.C. on land which was gracefully donated by Academos a school of higher education which in memory of the benefactor was named the Academy (Devlin, 1998).

Plato until his death was an active scholar and the president of the Academy of Athens which was an institution dedicated to research and teaching of philosophy and the sciences and is regarded by scholars as the spiritual fount of modern Universities.

In fact, such was the attachment of the Helene scholars to the knowledge of mathematics and geometry that it is said that an epigraph was permanently displayed at the entrance of the Academy stating: whoever has not mastered geometry may not enter.

Thus the modern day concept of academic refers to something or someone devoted to scientific research and teaching and is attributed to Plato's Academy who pioneered this idea some 2400 years ago.

Plato indicated that there was a third state beyond true and false where the two antithetical concepts are intertwined. Interestingly enough one of his students in the Academy of Athens was Aristotle. Another very important syllogistic contributor and Helene philosopher who unfortunately is rarely mentioned in the scientific literature is Zeno of Citium who lived approximately around 300 B.C. and was the founder of the Stoic school of logic (Devlin, 1998).

Zeno and the Stoics introduced most of the fundamental notions of contemporary propositional logic such as the patterns of reasoning which in contemporary logic actually are the patterns of interconnection between propositions that are completely independent of the operations of those propositions.

Another noteworthy fact is that the Stoics elucidated their logic not with algebraic notation, in other words, with letters indicating arbitrary propositions and symbols signifying connectives such as the ones used by contemporary logic but instead Stoic logic was simply expressed with written ancient Hellenic natural language.

The Stoics were such strong believers of their logic philosophy that they would follow to the letter logical reasoning even if that meant that they had to experience extreme hardship and greatly suffer by their actions which were prescribed by their formal

logic. Hence, the contemporary concept of stoical is attributed to this school of philosophy (Devlin, 1998).

Thus, the syllogisms of many earlier Helene philosophers culminated in Aristotle's formal definition of the cornerstone of syllogistic logic, the laws of logic also known as the laws of thought.

Many eons later Hegel, Marx and Engel contemplated and concurred with Plato's syllogisms. In the early nineteen hundreds Lukasiewicz first proposed a formal mathematical definition with an entire algebraic notation and axiomatic system of a triadic valued logic. This was the first alternative theory to Aristotle's binary logic, also known as bivalence, which was proposed over twenty three centuries earlier (Lukasiewicz, 1951).

In the world of mathematics Lukasiewicz is best known as the inventor of the parenthesis free notation, better known as Polish notation, most likely named after the country of origin of the inventor (Knuth, 1997). Lukasiewicz introduced a third logic value beyond the traditional true, logic 1 and false, logic 0 and identified this novel logic value as indeterminate or unknown which was assigned a numeric value ½ between true and false in his triadic valued logic.

Furthermore, Lukasiewicz researched quaternary and quinary valued logic eventually drawing the inference that multi-valued logic which is also known as multivalence could quite possibly be extended to a systematic infinite valued logic (Lukasiewicz, 1966).

Although the Aristotelian dyadic valued logic has been prevalent for twenty three centuries Lukasiewicz offered a very interesting alternative logic theory from which the derivative of infinite valued logic was a precursor to modern fuzzy logic. In 1965 Lotfi Zadeh published his seminal research paper entitled "fuzzy sets" which analysed the formal mathematical definition, properties and operations of fuzzy set theory and by extension fuzzy logic (Zadeh, 1965).

Fuzzy set theory described an element membership function taking any possible value between true 1 and false 0 in other words operating over the infinite space interval of real numbers [0.0, 1.0] which in principle is very similar to multi *n* valued logic which can take values from two to infinity.

When *n* takes a value of two according to the classical dyadic logic theory all the laws of classic Aristotelian binary valued logic hold true. However, the principle difference between multi *n* valued and fuzzy logic is that the truth values in the latter are infinite and bound by the inclusive interval [0.0, 1.0] which is identical in form to the original fuzzy set theory proposed by Lotfi Zadeh (Zadeh, 1965).

2.4 Developments in fuzzy information retrieval

Information retrieval techniques generally represent documents and queries through surrogate sets of key words. Considering that this strategy reflects only partially a piece of the semantic exegeses of the corresponding documents which the surrogate sets represent then the matching of a document to a query can only be an approximation.

In terms of fuzzy logic this approximation and obscurity can be modelled by considering that each query elucidates a fuzzy set and that each document has a degree of membership to that set. However, the foundation for the majority of both relational database and traditional information retrieval systems has long been Boolean logic.

For instance, when Boolean logic is used in an information retrieval system the query key words are connected by the very well known logical operators *AND*, *OR* and *NOT* and after the query is issued the information retrieval system returns all the documents which contain combinations of the supplied query key words satisfying the constraints of the specified logical Boolean operators.

Often in numerous information retrieval systems auxiliary means are made available to permit proximity and truncation searching. When proximity, truncation and field restricted searching are combined with databases connected to the internet the user is allowed to construct and issue remotely quite sophisticated queries (Hartley, Keen, Large and Tedd, 1990).

The Boolean logic model is very well accepted and also well understood but bears inherent limitations which make it less attractive especially for inexperienced users of information retrieval systems who without extensive exercise and practice find it difficult to construct any but the most uncomplicated queries constrained by Boolean logic operators (Cooper, 1988).

Consequently, experienced users frequently must interfere to issue a search on behalf of the novice user who feels the actual information need. Furthermore, the variable length of the query results and the poor precision of the system might force a series of information retrieval episodes. Under such conditions it would be a priori impossible to predict how many recurrences of the information retrieval episodes would be required to satisfy the initial information need.

Another important limitation of Boolean information retrieval systems is that often the results are partitioned into two discrete categories, namely the documents that are equivalent to the logic query constrains and the rest which belong to the complement of the previous set (Cooper, 1988). Thus, the entirety of the retrieved documents is alleged to be of equal importance to the information seeker and often there is no facility by which documents can be positioned in order of descending grade of relevance.

Ultimately, in the Boolean logic model there are no apparent facilities by which an information seeker can attribute a grade of importance to the various query key words since all key words implicitly assume the weight of unity if they are present or the weight of zero if they are absent in accordance to Boolean logic.

These Boolean logic information retrieval systems limitations have given researchers the opportunity for the development of fuzzy set models which attracted much attention in the early nineteen eighties due to their ability to naturally accept progressive grades of membership in contrast with the strict binary membership enforced by Boolean logic (Bookstein, 1986).

In addition, Van Rijsbergen has suggested that the foremost duty of an information retrieval system is the discovery of specific documents d from a large collection of documents in reaction to a query q. These together satisfy the logical implication $d \rightarrow q$. The symbol $\rightarrow$ symbolizes the kind of implication which is elucidated by the distinct logic that is espoused between d and q which are formal symbols of the semantics of the document and the query (Van Rijsbergen, 1986).

Van Rijsbergen's point of view is apparently a long way from what practical every day users think of information retrieval systems; however, it is not difficult at all to deduce possibly all existing information retrieval models from Van Rijsbergen's a priori and all encompassing elucidation.

For instance the Boolean and probabilistic models can be expressed as variant forms of logical implication and already this strategy has been investigated in depth most notably in the field of inference networks. On the other hand, weighted Boolean queries facilitating information retrieval models with the ability to produce unambiguous results would not be impossible to create with the use of fuzzy sets as suggested by Bookstein (Bookstein, 1980).

In such a model for instance the weight of key word T_k in document D_j would symbolize the degree of membership of the document D_j belonging to a set of documents indexed by the key word T_k and if the weight was equal to 1 then document D_j would belong completely to the set of documents indexed by T_k a

weight of zero would denote D_j absence from the set and an intermediate weight would symbolize a progressive degree of membership.

In a similar manner the impact of synthesized Boolean statements can also be expressed via the use of the fuzzy set theory (Bookstein, 1980). Salton suggested that it is possible to expand the global matching operations in effect under the vector document model to include Boolean query construction ensuring that dynamic query modification methods such as relevance feedback could have a bearing on Boolean query creation.

Salton's a priori elucidation of the expanded Boolean logic is an all encompassing theory embracing equally the vector document model and the Boolean query model and treating them as special cases of the generalised theory (Salton, 1989).

As has been mentioned earlier, a dyad of operands connected together via a correlating logical operator such as *AND*, *OR* and *NOT* constitutes a Boolean statement conveying key word associations such as synonym associations for *OR*, key word phrase associations for *AND*, and antithesis associations for *NOT*.

Furthermore, Buell suggested a dyadic stage method for every dyad of a Boolean query Q and a document D which in principle could estimate the query document resemblance $sim(D,Q)$. In the first instance every query key word q_i in Q is substituted by a function $F(D,q_i)$ which in the traditional Boolean model takes the value of 1 if key word q_i is present in D and zero if key word q_i is absent.

In the case of multiple Boolean operators in excess of a dyad the evaluation process executes in a recursive manner processing one expression at a time beginning with the innermost clause and working itself towards the outer layers (Buell, 1981).

For instance if a query Q is formulated such as $((\alpha \; or \; \beta) \; and \; not \; \gamma)$ and a document D includes the key word α but not key word β or key word γ, the function $F(D,q_i)$ is then computed as follows $F(D,\alpha)=1$ and $F(D,\beta)=F(D,\gamma)=0$. The initial function $F(D,\alpha \; or \; \beta)$ then is estimated to 1 and the entire equation results to $1\cdot(1-0)=1$, inferring that the value of D with respect to Q is one which symbolizes a match.

Over the years various proposals have appeared for expanding the traditional Boolean model the most significant being the initiation of progressive document key word weights indicating the magnitude of the impact of the discrete key words to the documents of a collection. For instance, if the weight of a query key word q_i in a document D is equivalent to k then the function $F(D,q_i)$ is equal to k, where k may be assigned any possible value in the interval from 0 to 1.

Furthermore, if the reader considers a fuzzy sets environment then the function $F(D,q_i)$ can be identified as the membership function of the key word q_i signifying the degree of q_i belonging to the set of documents indexed by q_i with the computation equations of the aforementioned analysis being identical, however the values that the formulae can now take are able to vary progressively in the interval between zero and one.

Hence, the fuzzy set environment is similar in behaviour and reduces to a pure Boolean model when the document key word weights are limited to the dyadic values of zero and one. For instance, if the reader considers the previous sample query $((\alpha \text{ or } \beta) \text{ and not } \gamma)$ and assumes that $F(D,\alpha) = 0.6$, $F(D,\beta) = 0.2$, and $F(D,\gamma) = 0.1$. Hence, $F(D,\alpha \text{ or } \beta)$ evaluates to $max(F(D,\alpha),F(D,\beta)) = 0.6$ and the final formula result value then is $0.6 \cdot (1 - 0.1) = 0.54$.

Thus, the fuzzy set methodology is much less limiting than the pure Boolean model for the reason that it makes available progressively graded retrieval results in descending order of query document resemblance while maintaining the structured query syntax of the typical Boolean environment.

The positioning facility is also able to limit the length of the retrieval results by recovering in each occasion a limited number of the top rated documents as prescribed by individual user wishes (Salton, 1989). Another interesting strategy as suggested by Kamel et al. is to search for fuzzy descriptors which are either quantitative or qualitative and at the same time they are indicative of the value of the document information (Kamel, Hadfield and Ismail, 1990).

For instance, the qualitative fuzzy descriptors encompass equally numerical values such as small, large, short, and tall and linguistic concepts such as polite and colourful whereas antithetically the fuzzy quantitative descriptors encompass values such as few, many, at least, and most of.

The utilisation of a key word in a document may as well be illustrated with a fuzzy method such as for instance moderately essential or quite imperative and the product of the retrieval could be ranked as highly relevant or partially relevant in which case the dilemma becomes how to convert such linguistic key words into the membership function values connected with fuzzy information retrieval.

In addition, a fuzzy matching procedure can also be synthesized together with additional matching procedures such as for instance a weighted Boolean process (Bordogna and Pasi, 1993). Another interesting approach as suggested by Kowalski is to utilize the fuzzy logic model reflecting non-specificity in an information retrieval system in order to provide to the user the benefit of automatic correction of misspellings or variations of the same stem key word (Kowalski, 1997).

For instance if the reader considers exercising an information retrieval system with the facility of a fuzzy query instrument upon a random document collection and by accident misspells the key word cranberry as cranbeery the information retrieval system would in principle proceed with one of two courses of action.

Firstly, it would indicate the misspelled key word cranbeery and propose from a well known thesaurus such as Oxford's the correct key word spelling cranberry or perhaps even synonyms such as crimson, ruby, burgundy and request the user's feedback on which key word should be utilized in the final query.

Secondly, an alternate course of action would be that the fuzzy information retrieval system upon detection of the misspelled key word cranbeery would immediately

without any user interaction engage a well accepted stemming algorithm such as Porter's and automatically enter the stem root key word cranb in the final query to be issued which is hoped would return documents relative to cranberries as the root is identical for both key words (Porter, 1997).

In such an information retrieval system the user entered queries are automatically stemmed and the retrieval results equally encompass documents which are quite closely related to the stem key words or actually include them. Another interesting point of view upon fuzzy queries is with regard to the collections of the retrieval results.

This view is analogous to typical information retrieval systems where by the retrieval results are positioned in order of relevancy. However the antithesis with regard to the fuzzy information retrieval systems is that the threshold of relevancy is increased to encompass further documents which may be of interest to the user and otherwise would have gone unnoticed (Korfhage, 1997).

Another interesting approach is to espouse a thesaurus in the information retrieval system. As it has been mentioned earlier the thesaurus is used in order to expand the collection of index key words in the query automatically or interactively with related or synonym key words. The system uses key words from the thesaurus in order to discover further useful and relevant documents beyond the usual results retrieved by a typical user composed query.

In addition, a thesaurus can also be combined with fuzzy sets in order to improve the modelling of a fuzzy information retrieval system. For instance, a thesaurus could be created using a key word to key word connection matrix c in which all the rows and columns are related to the index key words of the document collection (Ogawa, Morita and Kobayashi, 1991).

Thus, in such a key word connection matrix c, a normalized connection parameter $c_{i,l}$ between a dyad of key words k_i and k_l could be defined by the following equation.

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}} \tag{2.1}$$

Where n_i is the amount of documents which include the k_i key word, n_l is the quantity of documents which include the k_l key word and $n_{i,l}$ is the number of documents which include both aforementioned key words. Furthermore, such a key word connection matrix has often been combined with clustering algorithms equally for local and global document analysis (Xu and Croft, 1996).

A fuzzy subset A of a universe of discourse U is characterized by a membership function which associates with each element u of U a number $\mu_A(u)$ in the interval [0, 1]. Let B be a fuzzy subset of U and $\overline{A}$ be the complement of A relative to U.

$$\mu_{\overline{A}}(u) = 1 - \mu_A(u) \tag{2.2}$$

$$\mu_{A \cup B}(u) = max(\mu_A(u), \mu_B(u)) \tag{2.3}$$

$$\mu_{A \cap B}(u) = min(\mu_A(u), \mu_B(u)) \tag{2.4}$$

Another interesting approach is to combine the key word connection matrix c with fuzzy sets associated to individual index key words k_i in such a way that for each individual fuzzy set a document d_j can be attributed a degree of membership $\mu_{i,j}$ computed as follows.

$$\mu_{i,j} = 1 - \prod_{k_l \in d_j} \left(1 - c_{i,l}\right) \tag{2.5}$$

The membership equation computes an algebraic subtraction of one minus the product of one minus $c_{i,l}$, over all key words in the d_j document. Hence, the connection matrix c defines a fuzzy set associated to each index term k_i and each document d_j has a membership $\mu_{i,j}$ to that fuzzy set. If document d_j contains terms connected to term k_i then d_j belongs to the fuzzy set. If $c_{i,l}$~1 at least one term of d_j is strongly connected to term k_i and then $\mu_{i,j}$~1. If all terms of d_j are loosely connected to term k_i then $\mu_{i,j}$~0 (Baeza-Yates and Ribeiro-Neto, 1999).

Furthermore, if a document d_j is a member of the fuzzy set correlated to the k_i key word then the key words contained in document d_j are connected in some way with key word k_i. In addition if as a minimum there is a single index key word k_l of document d_j which has a close connection to the index key word k_i in other words $c_{i,l} \cong 1$ then $\mu_{i,j} \cong 1$ and the key word k_i would be an appropriate fuzzy index for the d_j document.

In the occasion where all index keywords of document d_j are only remotely connected to k_i the index key word k_i is not an appropriate fuzzy index for document

d_j in other words $\mu_{i,j} \cong 0$ and antithetically to expectations the use of an algebraic sum of all key words in document d_j instead of the typical max operation permits a fluid conversion for the values of the $\mu_{i,j}$ membership function.

For instance, if cc_i is a pointer to the element i then a query in disjunctive normal form is $\vec{q}_{dnf} = cc_1 \vee cc_2 \vee \ldots \vee cc_p$ where p is the quantity of elements of $\vec{q}_{dnf}$ which follows a similar discovery process for finding relevant documents which is quite analogous to the classic Boolean procedure (Baeza-Yates and Ribeiro-Neto, 1999).

However the only antithesis with the classic Boolean procedure in this case is that the sets are fuzzy instead of crisp Boolean sets. For instance if the reader considers the fuzzy set D_α of all the documents d_j related to the index key word k_α with a degree of membership $\mu_{\alpha,j}$ greater than a prearranged threshold K then $\overline{D}_\alpha$ is the complement of the fuzzy set D_α.

Also, the fuzzy set $\overline{D}_\alpha$ is related to $\overline{k}_\alpha$ the complement of the index key word k_α and analogously fuzzy sets D_β and D_γ can be defined which are related to the index key words k_β and k_γ correspondingly. In addition due to the fuzziness of all the sets even in the occasion when a document d_j does not explicitly include the index key word k_α, but is somehow related to the concept conveyed by it, then d_j could still be a member of the fuzzy set D_α.

In addition, Ogawa et al suggested a method to encompass user relevance feedback into the aforementioned model (Ogawa, Morita and Kobayashi, 1991). More recently, Baranyi and Koczy proposed a very general neural network approach which computes various arbitrary kinds of key word weighting functions such as follows.

$$y_{l+1,j} = \sum_i f_{l,j,i}\left(y_{l,i}\right) \tag{2.6}$$

The aforementioned equation elucidates a very general type neural network where $y_{l+1,j}$ is the result of layer $l+1$, $j = 1,2,\ldots,n_{l+1}$, n_{l+1} is the quantity of neurons existing at layer $l+1$, $f_{l,j,i}\left(y_{l,i}\right)$ are all the key word weighting functions in a matrix form elucidating the connections between a dyad of layers l and $l+1$ (Baranyi and Koczy, 2001).

However, the quite expanded variety of arbitrary unfamiliar key word weighting functions puts forth immense computational resources and effort demands which may lead the computing system to terra incognita.

A common and well known problem among neural network researchers is that when training or searching for unknown functions especially if a quite ambitious approximation is sought this may lead to a very complicated mathematical and even perhaps unsolvable Gordian knot computational problem.

One possible course of action to avoid the dead end situation in this particular case is to substitute the unknown arbitrary key word weighting functions with linearly combined already known functions where only the simpler linear combination must

be trained thus pragmatically resulting into the summation of parallel layers which form the system output.

Further experiments suggest that the more fuzzy sets on each universe are used the more computing estimation is improved, however, the increased training parameters require more calculation time and effort so for optimal results a balance must be maintained between the two extremes (Baranyi and Koczy, 2001).

Another interesting approach was proposed by Bordogna and Pasi who elucidated a fuzzy linguistic approach to document retrieval by generalizing Boolean information retrieval which permits forming an appropriate environment for content choice based on searchers' queries (Bordogna and Pasi, 1993). As it has been suggested earlier by Yager and Kacprzyk weighted averaging operators elucidated in connection with a weighting vector are able to model a linguistic quantifier such as most of, at least, or a few (Yager and Kacprzyk, 1997).

Due to the fact that ordered weighted averaging operators are median operators and their behaviour lies between the logical *AND* which could be viewed as a min and the logical *OR* which could be viewed as a max the degree of deviation of the operator illustrates its proximity to the logical *OR* behaviour and could be elucidated as follows.

$$orness(\vec{W}) = \left(\frac{1}{n-1}\right)\sum_{j=1}^{n}\left((n-j)\cdot w_j\right) \tag{2.7}$$

Where $\vec{W}$ is a weighting vector elucidated as $\vec{W} = [w_1, w_2, \ldots, w_n]$ such that $\sum_{j=1}^{n} w_j = 1$ with $w_j \in [0,1]$. Evidently, the principal limitation of the proposed approach is that

the documents must share the same logical structure consisting of a finite set of sections such as title, authors, abstract, body text and references signifying that this approach is only valid for homogeneous documents (Bordogna and Pasi, 2001).

Fuzzy methodologies have often been used for document clustering taxonomies. A very popular fuzzy clustering algorithm is fuzzy hierarchical clustering which is also known as agglomerative hierarchical clustering (Rasmussen, 1992; Salton, 1989). Initially the agglomerative hierarchical clustering algorithm treats every document as an individual cluster and then it recursively fuses the most analogous dyad into a single cluster continuing until the similarity between any dyad of clusters falls below a predefined heuristic threshold.

There are numerous ways to measure the similarity between any dyad of clusters and the agglomerative hierarchical clustering algorithm is not restrictive at all on the selection of a similarity measure. The most well known approaches are notably Salton's document vector cosine similarity measure and also the complete link clustering whereby either the minimum or the maximum or the average of the similarities is measured between any dyad of objects each from one cluster (Salton, Wong and Yang, 1997; Salton, Allan, Buckley and Singhal, 1997).

Another very popular fuzzy clustering methodology is the fuzzy c-means clustering which is a family of algorithms forming document taxonomies in an iterative manner by minimizing an objective function (Bezdek, 1980; Bezdek, Hathaway, Sabin and Tucker, 1987).

Another interesting approach proposed by Kraft and Chen was the use of fuzzy clustering algorithms to juxtapose document classifications. In order to measure differences between hierarchical and c-means fuzzy document clustering hardening was performed to the fuzzy document clusters acquired by the latter, taxonomizing every document to precisely a single cluster in which the membership function attained the maximum value over all other clusters (Kraft and Chen, 2001).

Furthermore a strategy to derive fuzzy rules from the document taxonomies is the following. For instance, if in either a document or a query a key word t_i has a weight w_i which is equal to or higher than its threshold W_i then in the same document or query a closely associated key word t_j has a weight w_j which is equal to or higher than its threshold W_j thus,

$$[w_i \geq W_i] \rightarrow [w_j \geq W_j] \tag{2.8}$$

$$[w_j \geq W_j] \rightarrow [w_i \geq W_i] \tag{2.9}$$

Therefore, under this approach in every document or query every key word has a closely associated counterpart key word so all of the key words in the same document or query are treated as dyads.

Furthermore a normalisation of the document clusters centres vectors takes place whereby all the weights for every key word are divided by the maximum weight for that key word in any document. Hence, if for instance the key word weights were chosen to be frequencies this method is an easy way to normalise all the key word weights in order to take values in the interval [0, 1].

Also according to Kraft and Chen's methodology the list of key words for every cluster centre was sorted in decreasing order of key word weights, key word dyads were formed from the top two or four key words in the list and if multiple instances of an identical dyad were found in more than one document cluster centre then the key word dyads were fused by choosing the minimal weight for every key word over all instances.

Thus for all key word dyads the aforementioned fuzzy rules hold equally true elucidating that the appearance of key word t_i with a weight w_i of at a minimum W_i always goes together with the appearance of key word t_j with a weight w_j of at a minimum W_j and vice versa.

Furthermore a strategy to perform query modification is the following. For instance if the reader considers a query and a fuzzy rule of the following forms,

$$q = < w_{q1}, w_{q2}, \ldots, w_{qs} > \tag{2.10}$$

$$[w_i \geq W_i] \rightarrow [w_j \geq W_j] \tag{2.11}$$

and also if the reader takes into account that $w_{qi} \geq W_i$, however, $W_j \geq w_{qj}$ for key words t_i and t_j in the query q then the effect of the fuzzy rule to q will yield $q' = < w'_{q1}, w'_{q2}, \ldots, w'_{qs} >$ where $w_{qi} = w'_{qi}$ for every $i \neq j$, but where $w'_{qj} = W_j$ the fuzzy rule coerces the associated key word to take a greater weight in the query equal to its threshold.

Hence, the proposed fuzzy clustering methodology can be effectively used to taxonomise documents in clusters which can further infer fuzzy rules to capture semantic associations between index key words (Kraft and Chen, 2001).

In synopsis, there are numerous fuzzy information retrieval research strategies addressing various issues, such as common user mistakes like the misspelling of terms in queries, document memberships to fuzzy query sets, quantitative and qualitative fuzzy document descriptors, query fuzzy key-word descriptors, fuzzy integration of a thesaurus to expand on the available index key-words, fuzzy rules for term weighting, fuzzy document clustering, etc.

Although there are all these fuzzy information retrieval research strategies addressing various issues, none addresses the information overload problem, which is now more evident than ever. Enter a simple query in any commercial information retrieval system and millions of results ensue.

Finding the relevant information inside this gigantic answer set is an admirable achievement on its own. This work addresses this issue of identifying and bringing forward relevant documents in a gigantic answer set, where a user would normally give up. Also, this work investigates methods of improving the effectiveness of current information retrieval systems.

CHAPTER 3 – Bidirectional Fuzzy Logic

3.1 Introduction

Bidirectional, symmetrical fuzzy logic goes beyond classic logic, which only deals with true and false, i.e. {0, 1}. Zadeh conceived and developed a type of logic that can contain values in [0, 1] which is called fuzzy logic today (Zadeh, 1965).

On the other hand, many-valued logic was first studied by Lukasiewicz, who created a so-called triadic logic by adding a 1/2 into the binary set {0, 1}. Knuth modified Lukasiewicz's triadic logic to {-1, 0, +1} rather than Lukasiewicz's {0, 1/2, 1}.

This research synthesizes fuzzy logic with Knuth's {-1, 0, +1} triadic logic and focuses on developing a bidirectional fuzzy logic theory which can be applied to information retrieval.

3.2 Overview of Logic

Aristotle, a Hellene philosopher who lived during the fourth century BC, laid the foundations for modern mathematical logic with his syllogistic logic (Devlin, 1998). Lukasiewicz, while best known for his parenthesis free notation (Knuth, 1997), also known as Polish notation, suggested an alternative triadic logic in the early nineteen hundreds (Lukasiewicz, 1951).

In 1965, Zadeh published his promethean fuzzy sets paper which was the foundation of the theory of fuzzy logic (Zadeh, 1965; Kosko, 1999). Knuth also suggested a triadic logic quite analogous to Lukasiewicz's hoping that it would contribute even

more elegance to modern mathematics than the classical Aristotelian dyadic logic (Brule, 1986).

The principal distinction between the Lukasiewicz and the Knuth triadic logic was that the latter applied to the following triadic set {-1, 0, +1} rather than {0, 1/2, 1}, which was the domain of the former. There is additional evidence of research quests in the same direction, most notably in certainty theory.

Certainty theory is an extension of probability theory in continuous symmetrical intervals [-x, x], where $x \in [1, 10]$, and typically the full range of values, counting 0 between and including the symmetrical peaks x and –x, are commonly referred to as certainty factors (Hopgood, 2001).

In the early nineteen eighties several variations of certainty factors were extensively applied in the field of artificial intelligence and distinctively in the area of expert systems. One such system is *MYCIN* which is concerned with the medical diagnosis of infectious diseases. The successor of this system is called essential *MYCIN* also known as *EMYCIN* which is not restricted solely to medical diagnosis and incorporates a simplified uncertainty management process (Hopgood, 1993).

For instance, if the reader considers in *EMYCIN* a given hypothesis H then its certainty C(H) can take any possible values in the interval [-x, x]. The relationship among the certainty value C(H), the hypothesis H and the probability of the hypothesis P(H) is illustrated in the following Table 3.1.

Table 3.1. Certainty factors, hypotheses and probabilities correlations.

Certainty Values C(H)	Hypothesis H	Probability P(H)
C(H)=x	If H is known to be true	P(H)=1.0
C(H)=0	If H is unknown	P(H) remains at its prior value
C(H)=-x	If H is known to be false	P(H)=0.0

Thus, it is obvious that certainty theory is an extension of mathematical probabilities mapping the classical probabilities space as follows P(H): [0, 1] → C(H): [-x, x].

Furthermore, in information retrieval, the notion of statistics of negative scale can be quite instrumental. Therefore, to generalize and fuzzify Knuth's triadic valued logic into the interval [-1, 1] may prove valuable. In other words, we will develop a bidirectional, symmetrical fuzzy logic whereby the membership function values operate over the continuous interval range of real numbers [-1.0, 0.0] U [0.0, 1.0].

This bidirectional, symmetrical fuzzy logic has a natural application to three research areas of information retrieval, namely document-to-query or document-to-document similarity, relevance feedback, and query reformulation.

3.3 Information Retrieval Applications

In the first instance, when the area of concern is document-to-query or document-to-document resemblance a well known analogy measure is the *cosine document similarity formula* introduced by Salton, which operates in the *document vector space model* (Salton, 1989; Salton, Wong and Yang, 1997).

Although the natural cosine function is evaluated over the continuous cyclic angular displacement interval $[0, \ 2\pi]$, when it comes to information retrieval the cosine document similarity measure is typically restricted to the top rightmost quadrant $\left[0, \ \frac{\pi}{2}\right]$. This limited measurement space is due to the fact that all known document vector dimension evaluation approaches -also known as *term weights*- are normalized to take positive values.

Thus, the introduction of the bidirectional, symmetrical fuzzy logic space theory in this case allows document vectors to attain their full natural cyclic 360° angular displacement instead of the rather narrow 90° restricted space of the top rightmost quadrant of the traditional information retrieval document vector angular displacement.

In the second instance, when the area of concern is relevance feedback, the computer user who is the expert document reader is allowed by the information retrieval system to assess document relevance by assigning positive or negative impacts (weights) to prior retrieved documents. This weighting based on bidirectional symmetrical fuzzy logic permits the information retrieval system to potentially insert or remove terms from a prospective new automatically generated system query (Salton and McGill, 1997).

In the third instance, when the area of concern is query reformulation, the computer user who is the expert document reader could be allowed by the information retrieval

system to further assign positive or negative weights to the terms considered for incorporation into the new reformulated query (Salton and McGill, 1997).

3.4 Overview of Fuzzy Logic

As mentioned earlier Zadeh's fuzzy logic is characterized by a membership function lying in the positive unit interval. If Zadeh's definition is modified to include the additional negative space proposed by Knuth's triadic valued logic, then a fuzzy set A is characterized by a membership function mapping the elements of a domain, space, or universe of discourse X equally to both positive and negative unit intervals [-1.0, 0.0] $\cup$ [0.0, 1.0].

In this case, the set membership function is expressed as follows $\mu_A : X \rightarrow [-1.0, \quad 0.0] \cup [0.0, \quad 1.0]$. Then the value of $\mu_A(x)$ for the fuzzy set A is the membership value or the grade of membership of $x \in A$. Therefore, the element membership value illustrates the degree to which x belongs to the fuzzy set A, which can be either a discrete or alternatively a continuous function.

3.5 Preserved "Min-Max" Fuzzy Logic

Preserved "min-max" negative fuzzy logic is a model which upholds the standard unidirectional fuzzy set laws and expands the theory of fuzzy logic (Petratos and Chen, 2002). The basic idea behind this model is that by expanding the definition space of fuzzy logic the minimum and maximum membership values and their associations to the union and the intersection of fuzzy sets are preserved.

For instance, let U be a set such as $\mu_A : U \rightarrow [-1.0, \quad 1.0]$, which defines a fuzzy set. For each element x of U the union of fuzzy sets A and B, which is denoted by $A \cup B$, is defined by the following membership function $\mu_{A \cup B}(x) = \mu_A(x) \vee \mu_B(x)$, where $\mu_A(x) \vee \mu_B(x) = \max\{\mu_A(x), \mu_B(x)\}$.

Furthermore the intersection of a dyad of fuzzy sets A and B, denoted by $A \cap B$, is defined by the following membership function $\mu_{A \cap B}(x) = \mu_A(x) \wedge \mu_B(x)$, where $\mu_A(x) \wedge \mu_B(x) = \min\{\mu_A(x), \mu_B(x)\}$. In addition, the complement $\overline{A}$ of fuzzy set A is defined by the following membership function $\forall \quad \mu_A(x) \in [-1.0, \quad 1.0]$, then $\mu_{\overline{A}}(x) = -\mu_A(x)$, and therefore $\mu_{\overline{A}}(x) + \mu_A(x) = 0$ consequently.

3.6 Bidirectional Symmetrical Fuzzy Logic

An illustrative example may be instructive as to the nature of bidirectional, symmetrical fuzzy logic. Imagine there are three men, Good-Man, Neutral-Man and Bad-Man. Classical logic cannot represent this case since if "1" is good then "0" means "not good". "Not good" has two meanings: Neutral or Bad. In such a case, Knuth's logic is appropriate (Cignoli, D'Ottaviano and Mundici, 2000; Hajek, 1998).

A *bidirectional, symmetrical fuzzy logic* can be defined as a series of properties. For instance, let U be a set such as $\mu_A : U \rightarrow [-1.0, \quad 1.0]$ which defines a fuzzy set. For each element x of U the union of fuzzy sets A and B, denoted by $A \cup B$, is defined by the following membership function $\mu_{A \cup B}(x) = \mu_A(x) \vee \mu_B(x)$, where $\mu_A(x) \vee \mu_B(x) = \mu_A(x)$ if $|\mu_A(x)| > |\mu_B(x)|$.

Furthermore $\mu_A(x) \vee \mu_B(x) = \mu_B(x)$ if $|\mu_A(x)| < |\mu_B(x)|$. Also if $\mu_A(x) = \mu_B(x)$ a corollary is that $\mu_A(x) \vee \mu_B(x) = \mu_A(x)$. Furthermore, the intersection of fuzzy sets A and B, denoted by $A \cap B$, is defined by the following membership function $\mu_{A \cap B}(x) = \mu_A(x) \wedge \mu_B(x)$, where $\mu_A(x) \wedge \mu_B(x) = \mu_A(x)$, if $|\mu_A(x)| < |\mu_B(x)|$.

Furthermore $\mu_A(x) \wedge \mu_B(x) = \mu_B(x)$, if $|\mu_A(x)| > |\mu_B(x)|$. Also if $\mu_A(x) = \mu_B(x)$ then $\mu_A(x) \wedge \mu_B(x) = \mu_A(x)$. Let $\overline{A}$ be the complement of A relative to U then $|\mu_{\overline{A}}(x)| = 1 - |\mu_A(x)|$ where for simplicity if $\mu_A(x) > 0$ then $\mu_{\overline{A}}(x) > 0$ and if $\mu_A(x) < 0$ then $\mu_{\overline{A}}(x) < 0$.

In some cases, the bidirectional, symmetrical fuzzy logic can be restricted to a *unidirectional fuzzy logic*. Numerous laws hold true for both bidirectional fuzzy sets and *crisp sets* (i.e. sets which are defined exactly in classical set theory), and are listed below. We can easily see:

Property 1.

The idempotent law states: $A \cup A = A$, $A \cap A = A$ (3.1)

Property 2.

The commutative law states: $A \cup B = B \cup A$, $A \cap B = B \cap A$ (3.2)

Property 3.

The associative law states:

$$A \cup (B \cup C) = (A \cup B) \cup C, \; A \cap (B \cap C) = (A \cap B) \cap C \tag{3.3}$$

Property 4.

The distributive law states:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),\ A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \qquad (3.4)$$

Property 5.

The law of double negation states: $A = \overline{\overline{A}}$ (3.5)

Also, in general, there are properties which hold true for crisp sets, but do not hold true for fuzzy sets. For instance these two crisp-set laws do not hold true for fuzzy sets. The law of contradiction: $A \cap \overline{A} \neq \varnothing$. The law of excluded middle: $A \cup \overline{A} \neq X$. Here $\varnothing$ symbolizes an empty set and X the universe of discourse.

The equality principle of bidirectional fuzzy sets A and B is defined as follows: $A = B \Leftrightarrow \forall\, x, \mu_A(x) = \mu_B(x)$. In classical fuzzy set theory, there are two types of $\alpha-$cuts, strong and weak. Here another two types of $\alpha-$cuts are defined, which for the negative fuzzy space are named *inverse strong* and *weak* respectively. All these $\alpha-$cuts can be defined for symmetrical fuzzy sets A and B as follows:

Strong α–cut: $A_\alpha = \{x \mid\ |\mu_A(x)| > \alpha\ \}$, $\alpha \in [0,\ 1)$ (3.6)

Weak α–cut: $A_\alpha = \{x \mid\ |\mu_A(x)| \geq \alpha\ \}$, $\alpha \in (0,\ 1]$ (3.7)

Strong α_+–cut: $\{x \mid \mu_A(x) > \alpha\}$, $\alpha \in [0,\ 1)$ (3.8)

Weak α_+–cut: $\{x \mid \mu_A(x) \geq \alpha\}$, $\alpha \in (0,\ 1]$ (3.9)

Strong α_-–cut: $\{x \mid \mu_B(x) < \alpha\}$, $\alpha \in [0,\ -1)$ (3.10)

Weak α_-–cut: $\{x \mid \mu_B(x) \leq \alpha\}$, $\alpha \in (0,\ -1]$ (3.11)

The inclusion principle of bidirectional fuzzy sets A and B is defined as follows: $A \subseteq B \Leftrightarrow \forall_{\alpha \geq 0}(A_\alpha \subseteq B_\alpha)$. In other words, A is a subset of B if and only if the α-cut of A is a subset of the α-cut of B for every α defined.

3.7 Comparison of standard fuzzy logic and bidirectional fuzzy logic

As it has been stated earlier there are certain cases where the standard unidirectional fuzzy logic does not provide an adequate portrayal or rather produces an incomplete and perhaps confusing representation of events, however, bidirectional fuzzy logic in a more natural manner illustrates the facts and provides certain advantages over the standard unidirectional fuzzy logic as well.

More specifically, there are a few clear examples following this brief introduction in different scientific areas of physics and computer science which are firstly implemented with the standard unidirectional fuzzy logic and secondly are demonstrated with the bidirectional fuzzy logic to illustrate the differences between the two theories.

Example 1: In physics and distinctively in electricity, the alternating current or voltage alters current direction or voltage polarity as time progresses according to its respective angular displacement. If the angular displacement of the AC sinusoidal waveform is $\theta \in [-2\pi, 2\pi]$, the period is $T = 2\pi$, the amplitude of the voltage, or current is α and the equation describing the AC waveform is $f(\theta) = \alpha \cdot \sin(\theta)$.

In order to normalize the waveform to accurately represent it with the standard unidirectional fuzzy logic the equation is modified as follows, $\alpha = \frac{v}{|v_{\max}|}$, or $\alpha = \frac{i}{|i_{\max}|}$, $f(\theta) = \frac{(\alpha \cdot \sin(\theta)+1)}{2}$ and the standard unidirectional fuzzy graph is illustrated in Fig. 3.1 on the left. In order to represent the AC waveform with the bidirectional fuzzy logic there is no modification required on the equation, just the amplitude is modified as follows, $\alpha = \frac{v}{|v_{\max}|}$, or $\alpha = \frac{i}{|i_{\max}|}$, $f(\theta) = \alpha \cdot \sin(\theta)$ and the bidirectional fuzzy graph is illustrated in Fig. 3.1 on the right.

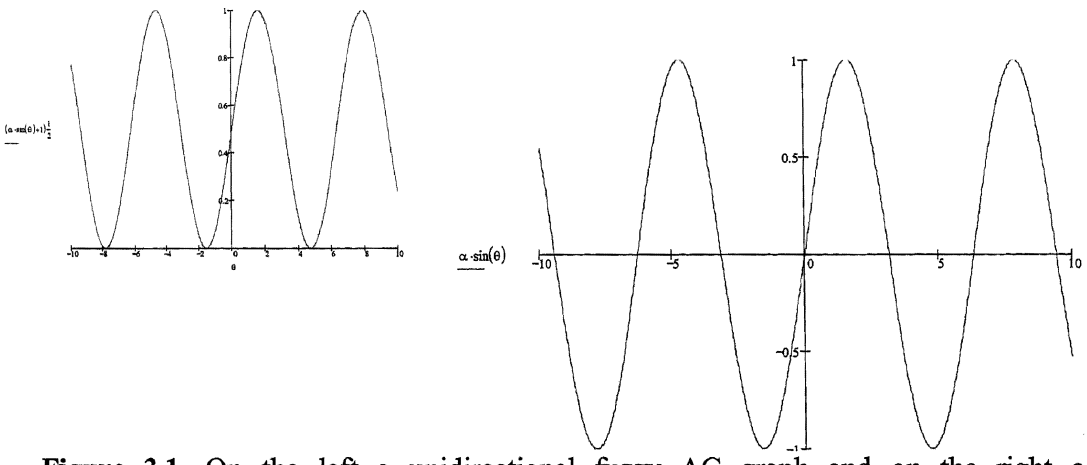

Figure 3.1. On the left a unidirectional fuzzy AC graph and on the right a bidirectional fuzzy AC graph.

Example 2: In Computer science and distinctively in information retrieval numerous scientists have relied on a very well known measure for document similarity, the cosine of the angle between two document vectors, which was first introduced by Salton (Salton, 1989).

Two documents are normalized so that they have the same number of dimensions. Each key term is a dimension represented numerically by the key word weight function and if a particular key word is absent in one or the other document then the key word weight assigned is zero.

The cosine equation describing the similarity of a dyad of document vectors is the following, $\cos(D_i, D_j) = \frac{\sum_{k=1}^{T}(t_{ik} \cdot t_{jk})}{\sqrt{\sum_{k=1}^{T}(t_{jk})^2 \cdot \sum_{k=1}^{T}(t_{jk})^2}}$, with an angle θ between the dyad of document vectors D_i, D_j where t_{ik}, t_{jk} are the corresponding terms in each document and T is the number of terms.

However, in traditional information retrieval the cosine document similarity measure is typically restricted to the quadrant $\left[0, \frac{\pi}{2}\right]$ due to the fact that all known document vector dimension evaluation approaches also known as key word weights are normalised to take values in one form or another of a positive continuous space interval including zero.

So, the introduction of the bidirectional fuzzy logic space theory in this case would allow document vectors to attain their full natural cyclic angular displacement $[0, 2\pi]$ instead of the rather narrow and restricted quadrant $\left[0, \frac{\pi}{2}\right]$ of the conventional document vector angular displacement in traditional information retrieval.

Therefore in order to normalize the restricted cosine function and represent it in the conventional unidirectional fuzzy logic space the equation should be modified as follows $\cos(D_i, D_j) = \frac{|\cos(\theta)| + \cos(\theta)}{2}$ and the standard unidirectional fuzzy graph is illustrated in Fig. 3.2 on the left. Antithetically, in order to represent the function with the bidirectional fuzzy logic there is no modification required on the equation, $\cos(D_i, D_j) = \cos(\theta)$ where θ is the angle between the two document vectors and the bidirectional fuzzy graph is illustrated below on the right.

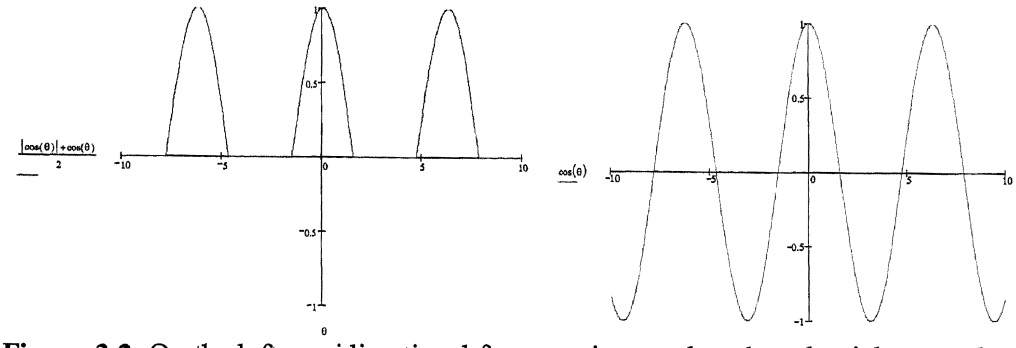

Figure 3.2. On the left a unidirectional fuzzy cosine graph and on the right a graph of a bidirectional fuzzy cosine function.

In the above graph the horizontal axis represents the angle θ in radians and the vertical axis represents the function $f(\theta)$ for the respective fuzzy representations.

3.8 Bidirectional fuzzy logic in the experimental information retrieval system

In order to investigate various aspects of this strategy the academic experimental information retrieval system called *ANACALYPSE* includes bidirectional fuzzy expert relevance feedback, document representation in the vector space model, and multiple weighting methodologies applied to the document retrieval sets.

Seen documents and feedback are combined into a single vector called *Metagram*, which is used for vector comparison with the unseen document vectors. For a detailed description of the procedure constructing the *Metagram* the reader is referred to please see section 6.6. The comparison of *Metagram* with an unseen document vector is accomplished through vector analysis and similarity computation of the cosine between this dyad of vectors.

All this information is stored and used to generate a knowledge base for further information retrieval sessions. The theoretical information retrieval background is the following. A document is represented by a term vector $D=(t_0,t_1,t_2,t_3,\ldots t_m)$, where t_x represents one of m terms which are content identifiers such as a word or a phrase from document D.

Terms can be reduced to their original root through a process better known as *word stemming* (Porter, 1997). Following the same philosophy, a query Q is represented by a term vector $Q=(q_0,q_1,q_2,q_3,\ldots q_k)$, where q_x represents one of k terms from query Q.

If w_{dm} or w_{qk} represent the weights of terms m, k in document D and query Q, respectively, then their term vectors can be normalized as follows. If term x is not present in D or Q, then $w_{dx}=0$, or $w_{qx}=0$, respectively. $D=(t_0,w_{d0};t_1,w_{d1};t_2,w_{d2};\ldots t_m,w_{dm})$, $Q=(q_0,w_{q0};q_1,w_{q1};q_2,w_{q2};\ldots q_m,w_{qm})$.

Furthermore a document collection N is $N = \{D_0, D_1, D_2, D_3, \ldots D_n\}$, which is represented by a matrix as follows.

$$N = \begin{array}{c} \\ D_0 \\ D_1 \\ D_2 \\ D_3 \\ \vdots \\ D_n \end{array} \begin{array}{c} \overbrace{\begin{array}{cccccc} t_0 & t_1 & t_2 & t_3 \ldots t_m & & \end{array}} \\ \begin{bmatrix} w_{00} & w_{01} & w_{02} & w_{03} & \cdots & w_{0m} \\ w_{10} & w_{11} & w_{12} & w_{13} & \cdots & w_{1m} \\ w_{20} & w_{21} & w_{22} & w_{23} & \cdots & w_{2m} \\ w_{30} & w_{31} & w_{32} & w_{33} & \cdots & w_{3m} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ w_{n0} & w_{n1} & w_{n2} & w_{n3} & \cdots & w_{nm} \end{bmatrix} \end{array} \tag{3.12}$$

The above collection consists of n documents with m terms and respectively m weights. The meaning of normalisation in this case is elucidated by the equal length of the vectors, as the document vectors all contain the same number m of terms and respectively m weights. Traditionally, the same term normalization and similarity functions apply between document vectors as between document and query vectors (Salton, 1989).

In synopsis, in this chapter an overview of classic logic and fuzzy logic is presented along with their corresponding applications to information retrieval. Herein the introduction of bidirectional fuzzy logic is presented as well.

Also, a comparison is offered of the standard uni-directional fuzzy logic and the new bi-directional fuzzy logic with descriptive examples both in physics and computer science.

More specifically, this presentation portrays the graphic differences of the previous standard uni-directional fuzzy logic and the new bi-directional fuzzy logic applications to information retrieval.

CHAPTER 4 – Research Methodology

4.1 Introduction

This chapter outlines the research methodology which is the systematic basis for the information retrieval experimentation reported in this thesis. The design of this methodology is intended to facilitate the comparison of a large number of similarity functions and term weights; as well as assessment of the effects of stop word removal, stemming, history of user web roaming, and finally comparison of the academic system of interest with a well accepted commercial information retrieval system.

4.2 Background

Previous information retrieval researchers have focused their studies on similarity heuristics and term weighting approaches in order to determine the most advantageous syntheses of formulae for optimal information retrieval systems performance (Salton and Buckley, 1988; Zobel and Moffat, 1998).

In this chapter a survey of similarity heuristics and term weighting strategies is presented and an experimental framework is proposed encompassing bidirectional fuzzy relevance feedback and the most advantageous components identified in a priori research studies.

Information retrieval research and more specifically its investigational aspects are greatly influenced by the specific environment under which experiments take place. A large number of factors and conditions such as the dynamic or static nature of the content and size of the database, the types of documents, the types of term weights,

the types of similarity heuristics, the types of evaluation measures, all play an important role in the information retrieval system overall performance.

However, although there are environmental conditions such as the dynamic nature of the database or the document types which are endemic to a specific environment and cannot be changed, the remaining influencing factors are usually a matter of human selection. Hence it is only natural to raise the question of which combinations from all the aforementioned components yield the optimum information retrieval system performance.

4.3 Similarity heuristics and term weighting approaches

The question of which syntheses of similarity heuristics and term weighting approaches yield optimum information retrieval system performance has been addressed by previous analysed experimental studies (Salton and Buckley, 1988; Zobel and Moffat, 1998).

The former study describes an experimental methodology of various term weighting formulae analysed into a six dimensional space *ddd-qqq* whilst keeping the same fixed similarity measure, notably the cosine. The latter study expands upon the former including additional components to reach an orthogonal eight dimensional space which allows numerous similarity measures which are specified as points in the eight dimensional space.

Hence this expanded experimental methodology permits the systematic and coherent exploration of the similarity space. Although this expanded eight dimensional space

experimental methodology is sufficiently general that most measures can be described in the same framework the current study expands upon it by including an additional new dimension due to the introduction of bidirectional fuzzy relevance feedback in the experimental framework to reach an expanded nine dimensional space.

Furthermore, although previous studies have focused on the similarities between queries and documents in the current study the emphasis of the experiments is on the similarities of unseen documents to an expanded compound information structure here termed a Metagram (Petratos, Chen, Wang & Forsyth, 2002).

The reason for this comparison shift is elucidated by the scope of this research. The primary aim of the current research is to propose methodologies of increased effectiveness for automatic information retrieval systems, with semantically relevant output compared to the subject matter expert's evaluation of relevance. Furthermore the experiments have two objectives:

- To automatically compute the relevance of a large number of unseen documents based on expert relevance feedback on a small number of evaluated documents.
- Moreover, to perform a systematic comparison of similarity heuristics and evaluate the effectiveness of the experimental information retrieval system *ANACALYPSE* compared to Google, the commercial information retrieval system compared to the expert's relevance standard.

Tables 4.1 to 4.5 include various components of simple statistical functions. For instance N is the number of documents, for each term t and each document d

containing t the frequency $f_{d,t}$ of t in d, for each term t the total number F_t of t in the collection, the number f_t of documents containing term t, $f_d = |d|$ is the number of term occurrences in d, the set T of distinct terms in the database, T_d in document d, T_μ in Metagram μ and $T_{d,\mu} = T_d \cap T_\mu$, the intersection of the sets of terms in the document and the Metagram.

Table 4.1. Similarity functions $S_{d,\mu}$.

ID	Description	Formulation
A	Inner product	$S_{d,\mu} = \sum_{t \in T_{d,\mu}} \left(w_{d,t} \cdot w_{\mu,t} \right)$
B	Cosine	$S_{d,\mu} = \frac{\sum_{t \in T_{d,\mu}} \left(w_{d,t} \cdot w_{\mu,t} \right)}{W_d \cdot W_\mu}$
C	Probabilistic	$S_{d,\mu} = \sum_{t \in T_{d,\mu}} \left(C + w_t \right)$
D	Alternative probabilistic	$S_{d,\mu} = \sum_{t \in T_{d,\mu}} \left(C + w_t \right) \cdot r_{d,t}$
E	Alternative inner product	$S_{d,\mu} = \sum_{t \in T_{d,\mu}} \frac{w_{d,t}}{W_d}$
F	Dice formulation	$S_{d,\mu} = \frac{2 \sum_{t \in T_{d,\mu}} \left(w_{d,t} \cdot w_{\mu,t} \right)}{W_d^2 + W_\mu^2}$
G	Jaccard formulation	$S_{d,\mu} = \frac{\sum_{t \in T_{d,\mu}} \left(w_{d,t} \cdot w_{\mu,t} \right)}{W_d^2 + W_\mu^2 - \sum_{t \in T_{d,\mu}} \left(w_{d,t} \cdot w_{\mu,t} \right)}$
H	Overlap formulation	$S_{d,\mu} = \frac{\sum_{t \in T_{d,\mu}} \left(w_{d,t} \cdot w_{\mu,t} \right)}{\min\left(W_d^2, W_\mu^2 \right)}$

In Table 4.1: C, D probabilistic measures contain a tuning constant C set to zero in (Zobel and Moffat, 1998). This Table 4.1 lists all the possible similarity functions used in the information retrieval experiments described in the latter part of this work.

Table 4.2. Term weights w_t (inverse document frequencies).

ID	Description	Formulation
A	Binary match	$w_t = \begin{cases} 1.0 & \text{if } t \in T_d \\ 0.0 & \text{otherwise} \end{cases}$
B	Logarithmic formulation	$w_t = \log\left(\frac{N}{f_t}\right)$
C	Hyperbolic formulation	$w_t = \frac{1}{f_t}$
D	Normalised formulation	$w_t = \log_e\left(1 + \frac{\max_{t \in T}(f_t)}{f_t}\right)$
E	Alternative normalised formulation	$w_t = \log\left(\frac{N - f_t}{f_t}\right)$
	Noise n_t of t	$n_t = \sum_{d \in D_t}\left(-\frac{f_{d,t}}{F_t}\log_2\frac{f_{d,t}}{F_t}\right)$
	Signal s_t of t	$s_t = \log_2(F_t - n_t)$
F	Signal formulation	$w_t = s_t$
G	Signal over noise	$w_t = \frac{s_t}{n_t}$
H	Noise adjusted formulation	$w_t = \max_{t \in T}(n_t) - n_t$
I	Entropy	$w_t = 1 - \frac{n_t}{\log_2 N}$

This Table 4.2 lists all the possible term weights, which are also known as inverse document frequencies and are used in the information retrieval experiments. The noise and the signal of a term t are defined in this table but have no ID associated with them as they are used in the latter F, G, H, I formulae defined herein as well (Zobel and Moffat, 1998).

Table 4.3. Document term weights $w_{d,t}$, $w_{\mu,t}$.

ID	Description	Formulation
A	Relative frequency formulation	$w_{d,t} = r_{d,t}$
B	Standard formulation	$w_{d,t} = r_{d,t} \cdot w_t$

This Table 4.3 lists all the possible relative document term weights, which are used in the information retrieval experiments. Both the relative frequency and the standard formulae utilize $r_{d,t}$ which is defined in the next Table 4.4 (Zobel and Moffat, 1998).

Table 4.4. Relative term frequencies $r_{d,t}$.

ID	Description	Formulation
A	Binary match	$r_{d,t} = \begin{cases} 1.0 & if\ t \in T_d \\ 0.0 & otherwise \end{cases}$
B	Standard formulation	$r_{d,t} = f_{d,t}$
C	Logarithmic formulation	$r_{d,t} = 1 + \log_e f_{d,t}$
D	Normalised formulation	$r_{d,t} = \frac{f_{d,t}}{\max_{t \in T_d}(f_{d,t})}$
E	Alternative normalised formulation	$r_{d,t} = K + K\frac{f_{d,t}}{\max_{t \in T_d}(f_{d,t})}$
F	Okapi formulation	$r_{d,t} = \frac{f_{d,t}}{f_{d,t} + W_d / Avg_{d \in D}(W_d)}$

This Table 4.4 lists all the possible relative term frequencies, which are used in the information retrieval experiments. In Table 4.4: E alternative normalised formulation contains a variable K which is a tuning constant with reported optimums 0.3 and 0.5 (Frakes & Baeza-Yates 1992, pg 370). The parameter K = 0.5 was used in (Zobel and Moffat, 1998).

Table 4.5. Document lengths W_d.

ID	Description	Formulation
A	Unit length	$W_d = 1$
B	Vector space formulation	$W_d = \sqrt{\sum_{t \in T_d} w_{d,t}^2}$
C	Approximate formulation	$W_d = \lvert T_d \rvert$
D	Alternative approximate formulation	$W_d = \sqrt{\lvert T_d \rvert}$
E	Alternative approximate formulation	$W_d = \log_2 \lvert T_d \rvert$
F	Byte size (Appellation due to alternate, $W_d = b_d$ which can be used, where b_d is the length of d in bytes.)	$W_d = f_d$
G	Alternative approximate formulation	$W_d = \sqrt{f_d}$
H-N	Pivoted method	$W_d = (1-s) + s \cdot \frac{W_d^{'}}{\alpha \upsilon_{d \in D} W_d^{'}}$

This Table 4.5 lists all the possible document lengths, which are used in the information retrieval experiments. Furthermore, in Table 4.5: F the byte size formulation alternatively can utilize $W_d = b_d$, where b_d is the length of document d in bytes.

In addition, in Table 4.5: H-N pivoted method $W_d^{'}$ is calculated using another length formulation such as method A to get method H, method B to get method I, method C to get method J and so forth. In the formulae H throughout N the variable s utilized is the slope and s = 0.7 was applied in the experiments reported by (Zobel and Moffat, 1998).

Table 4.6. Bidirectional fuzzy relevance feedback σ_d BDFRFB and history h_d adjusted term weight $w_{\sigma,t}$.

ID	Description	Formulation
N	BDFRFB adjusted weight	$w_{\sigma,t} = \sigma_d$
A	BDFRFB history adjusted weight	$w_{\sigma,t} = \frac{\sigma_d}{h_d}$
B	Alternative BDFRFB history adjusted weight	$w_{\sigma,t} = \frac{\sigma_d}{\sqrt{h_d}}$
C	Alternative BDFRFB history adjusted weight	$w_{\sigma,t} = \frac{\sigma_d}{\sqrt[3]{h_d}}$
D	Alternative BDFRFB history adjusted weight	$w_{\sigma,t} = \frac{\sigma_d}{\ln(h_d + 1)}$
E	Alternative BDFRFB history adjusted weight	$w_{\sigma,t} = \frac{\sigma_d}{\log(h_d + 1)}$

This Table 4.6 lists all the possible combinations of bidirectional symmetrical fuzzy relevance feedback as well as all the possible syntheses of history adjusted term weights, which are used in the information retrieval experiments.

Furthermore, in Table 4.6: σ_d is the expert's bidirectional symmetrical fuzzy relevance feedback for each examined document d and history h_d of document d is the number of steps taken from the root to where document d was found. Also, the ID: N stands for no history as steps are not considered in the formula.

One question is how documents should be weighted using this new information. Is there a single formula that should apply to all or are there more ways to weight them? What should be the criteria to satisfy?

For reasons subsequently explained two are the most important criteria:

- The formula should not change the sign i.e. not from + to - of the weight.
- The formula should monotonically decrease as the number of steps increases.

A few of the possible formulae are considered in the following graphs.

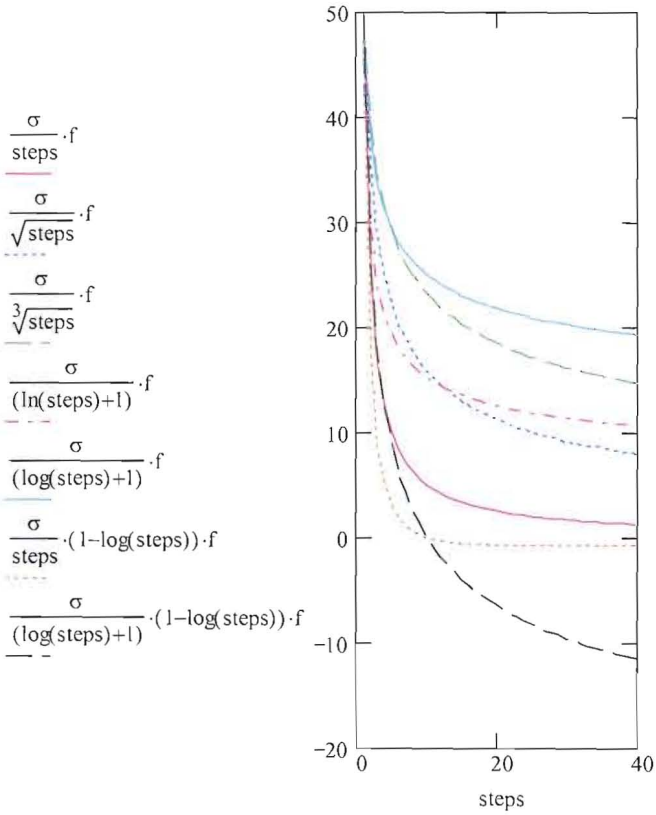

Figure 4.1. Graph of various history of information streams formulae $\sigma = 1$, f = 50, steps = 40.

The Figure 4.1 portrays all the possible combinations of positive bidirectional fuzzy symmetrical relevance feedback and history adjusted term weights to account for the

number of steps traversed from the root. In the above graph there is a number of different formulae considered for the optimum history adjusted term weighting strategy. The bidirectional fuzzy relevance feedback $\sigma = 1$, the term frequency of term x in document i is f = 50 and the history of document i is the number of steps taken from the root of the dendric structure to reach document i.

The rate of descent should be neither too harsh nor too slow. The sixth and seventh cases cross the steps axis to reach negative values in Figure 4.1 for the overall function affecting negatively the output and hence are unsuitable as they can confuse documents with a long history with documents belonging to the antithesis pole.

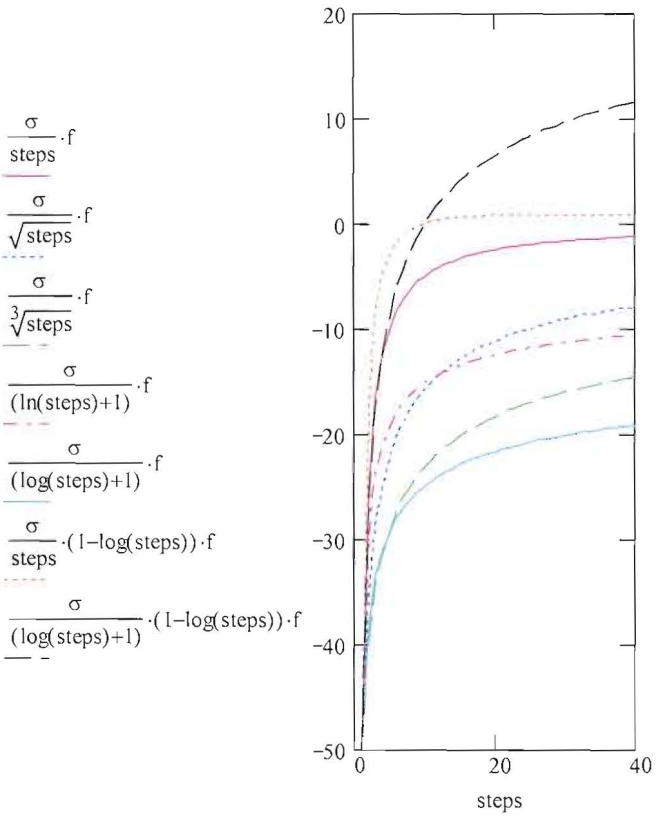

Figure 4.2. Graph of various history of information streams formulae $\sigma = -1$, f = 50, steps = 40.

The Figure 4.2 portrays all the possible combinations of negative bidirectional fuzzy symmetrical relevance feedback and history adjusted term weights to account for the number of steps traversed from the root.

The first formula divides the term frequency of term x in document i by the number of steps taken to reach document i while all the remaining formulae consider different logarithmic variants of the number of steps taken. The optimum function gradually converges to zero, that is it monotonically decreases anti-analogously as the number of steps increase.

The rate of ascent should be neither too harsh nor too slow. The sixth and seventh cases cross the steps axis to reach positive values in Figure 4.2 for the overall function affecting adversely the output and hence are unsuitable as they can confuse documents with a long history with documents belonging to the antithesis pole.

Therefore the most suitable functions for the purposes of the current research experiments are the first five as they all satisfy the required criteria. The same optimal characteristics are retained by the first five formulae as the bidirectional fuzzy relevance feedback enters the negative region of values.

Table 4.7. Putting it all together: example synthesis BAN-ABB-BBB in the new experimental nine dimensional space.

Component	ID	Formulation
Combining function	B	$S_{d,\mu} = \frac{\sum_{t \in T_{d,\mu}} \left(w_{d,t} \cdot w_{\mu,t} \right)}{W_d \cdot W_\mu}$
Weight of term t	A	$w_t = \begin{cases} 1.0 & if\ t \in T_d \\ 0.0 & otherwise \end{cases}$
BDFRFB adjusted weight of term t	N	$w_{\sigma,t} = \sigma_d$
Weight of term t in unseen document d	A	$w_{d,t} = r_{d,t}$
Relative frequency of term t in unseen document d	B	$r_{d,t} = f_{d,t}$
Weight of unseen document d	B	$W_d = \sqrt{\sum_{t \in T_d} w_{d,t}^2}$
Weight of term t in Metagram	B	$w_{\mu,t} = r_{\mu,t} \cdot w_{\sigma,t}$
Relative frequency of term t in Metagram	B	$r_{\mu,t} = f_{d,t}$
Weight of Metagram	B	$W_\mu = \sqrt{\sum_{t \in T_\mu} w_{d,t}^2}$

In Table 4.7 the selection of the individual components was reached after careful analysis (see next section of analysis) of the a priori experimental results (Salton & Buckley, 1988; Zobel and Moffat, 1998) according to the best performing combinations and furthermore based on the specific characteristics of the current experimental nature.

4.4 Analysis

After a careful analysis of ZM:(Zobel and Moffat, 1998) a few observations are possible. First, no single method of synthesis attains better performance than approximately two thirds of the optimum values that is 67% for recall-precision average.

Second, uniform poor performance characterises all the methods of syntheses with small divergence in performance results from one method to another and although the experimental framework of ZM contains all the components of SB:(Salton and Buckley, 1988) the experimental results of ZM cannot reproduce the experimental results of SB although they have been reproduced and validated by other researchers as well.

Third, although some form of the SB components exist in the ZM framework the exact original SB formulae do not exist in the ZM framework hence the SB experimental results cannot be reproduced in ZM.

Table 4.8. All SB components as they are mapped in ZM and the respective formulae differences.

SB ID	ID	Table	ZM Formula	SB Formula	Differ
x	A	T2	$w_t = \begin{cases} 1.0 & if\ t \in T_d \\ 0.0 & otherwise \end{cases}$	$w_t = \begin{cases} 1.0 & if\ t \in T_d \\ 0.0 & otherwise \end{cases}$	0
f	B	T2	$w_t = \log_e\left(1 + \frac{N}{f_t}\right)$	$w_t = \log\left(\frac{N}{f_t}\right)$	1
p	E	T2	$w_t = \log_e\left(\frac{N - f_t}{f_t}\right)$	$w_t = \log\left(\frac{N - f_t}{f_t}\right)$	1
b	A	T4	$r_{dt} = \begin{cases} 1.0 & if\ t \in T_d \\ 0.0 & otherwise \end{cases}$	$r_{dt} = \begin{cases} 1.0 & if\ t \in T_d \\ 0.0 & otherwise \end{cases}$	0
t	B	T4	$r_{d,t} = f_{d,t}$	$r_{d,t} = f_{d,t}$	0
n	E	T4	$r_{d,t} = K + (1 - K)\frac{f_{d,t}}{\max_{t \in T_d}(f_{d,t})}$	$r_{d,t} = K + K\frac{f_{d,t}}{\max_{t \in T_d}(f_{d,t})}$	1
x	A	T5	$W_d = 1$	$W_d = 1$	0
c	B	T5	$W_d = \sqrt{\sum_{t \in T_d} w_{d,t}^2}$	$W_d = \sqrt{\sum_{t \in T_d} w_{d,t}^2}$	0

The Table 4.8 lists all the SB information retrieval components as they are mapped in the ZM experimental framework. In Table 4.8: n the variable K = 0.5 in both SB, ZM hence the formulae compute identical values.

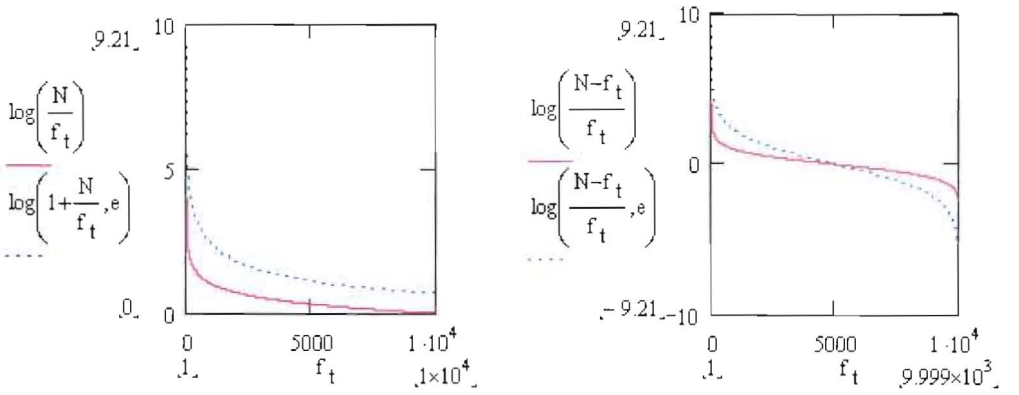

Figure 4.3. Components *f* left and *p* right as they are represented by SB top and ZM bottom.

The Figure 4.3 portrays the two information retrieval components, *f* on the left and *p* on the right as they are represented by SB, in the graph first row at the top and ZM, in the graph last row at the bottom. The two different representations of the two information retrieval components, *f* on the left, as well as *p* on the right by SB on the top and ZM on the bottom respectively, exhibit a great divergence in their output values.

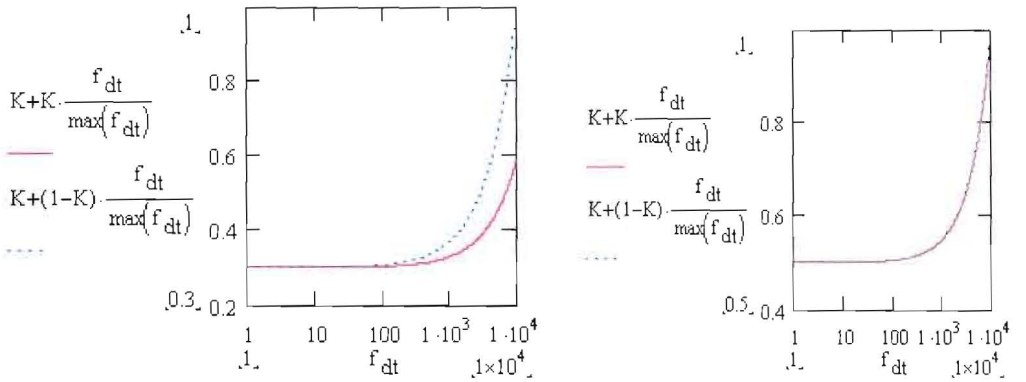

Figure 4.4. Component *n* with K=0.3 left and K=0.5 right as they are represented by SB top and ZM bottom.

The Figure 4.4 portrays a single information retrieval component, *n* on the left with K=0.3 and *n* on the right with K=0.5 as they are represented by SB, in the graph first row at the top and ZM, in the graph last row at the bottom. The two different representations of the *n* information retrieval component with K=0.3 on the left by SB on the top and ZM on the bottom respectively, exhibit a divergence in their output values.

In all the graphs the differences of the SB and ZM formulae are portrayed. Only component *n* yields the same values if K = 0.5 otherwise it diverges. Hence as the one is really a different study from the other the ZM experimental results cannot be compared with the SB experimental results which have been reproduced and validated by other studies as well (Sparck Jones and Willett, 1997). Therefore in the current study the SB experimental results are followed and the ZM framework is used but only with the SB original formulae.

Table 4.9. Experimental SB methods and results as they are mapped in the ZM framework.

SB Method	T1	T2	T3	T4	T5	T3	T4	T5	C1	C2	C3	C4	C5
tfc · nfx	B	B f	B	B t	B c	B	E n	A x	0.3630	0.2189	0.3841	0.2626	0.5628
txc · nfx	B	B f	A	B t	B c	B	E n	A x	0.3252	0.2189	0.3950	0.2626	0.5542
tfx · tfx	B	B f	B	B t	A x	B	B t	A x	0.3248	0.2166	0.2991	0.2365	0.5177
nxx · bpx	B	E p	A	E n	A x	B	A b	A x	0.3090	0.1441	0.3899	0.2093	0.5449
bfx · bfx	B	B f	B	A b	A x	B	A b	A x	0.2535	0.1410	0.3184	0.1781	0.5062
bxx · bpx	B	E p	A	A b	A x	B	A b	A x	0.2376	0.1233	0.3266	0.1563	0.5116
txc · txx	B	A x	A	B t	B c	A	B t	A x	0.2102	0.1539	0.3408	0.1620	0.4641
bxx · bxx	B	A x	A	A b	A x	A	A b	A x	0.1848	0.1033	0.2414	0.0944	0.4132

An implicit mapping rule deduced from Table 4.9 is that if there is an x in the middle of *ddd* or *qqq* of the SB method then the corresponding ZM framework T3 value is equal to A. Also, in Table 4.9 C1-C5 are five different document collections listing the corresponding average precision results for each SB method.

Furthermore, SB methods one and two are best in the group producing good comparable performances for all document collections. In the current study the experiments are on scientific text as all the experts are scientist researchers. In Table 4.9 collections C3 and C4 are scientific and technical document collections and the best SB method for these collections is method two.

Hence, the SB method that best suits the current study is method two $txc \cdot nfx$. Another important difference in the current study is that the methods are expressed in

terms of document and Metagram $ddd \cdot \mu\mu\mu$ instead of document and query $ddd \cdot qqq$ as in the SB study.

The Metagram is a data structure which contains all the terms from a large number of documents hence its size and contents are more like a document rather than a simple, small query. Therefore according to the SB experimental results the method that suits best the current study is $txc \cdot txc =$ BA-ABB-ABB.

However in the current study there is also the additional new dimension of the adjusted term weight for the Metagram to consider. Hence the new synthesis of the experimental method becomes BA*N*-ABB-BBB (see Table 4.7), where the third component *N* is the new adjusted term weight and in the last triad the first component B takes into account the new adjusted term weight. All the following possible combinations [ABFGH]A[NABCDE]-ABB-BBB are considered in the experimental phase.

In synopsis, in this chapter an overview of the background of information retrieval methodologies is presented. Previous information retrieval research studies have focused on similarity heuristics and term weighting approaches in order to determine the most advantageous syntheses of formulae for optimal information retrieval systems performance (Salton and Buckley, 1988; Zobel and Moffat, 1998).

Herein a comparative analysis is presented of these research methodologies and their respective results indicating differences in the formulae used and the ensuing differences of previous experimental results.

The current study utilizes the best performing information retrieval components from the previous studies and introduces bidirectional symmetrical fuzzy relevance feedback and history adjusted term weights to account for the number of steps traversed from the root. The current study attempts a plethora of combinations of these formulae which are used in the current information retrieval experiments.

CHAPTER 5 – Information retrieval metrics and experimentation

5.1 Organisation

The information presented in Chapter 5 and the topics discussed in Chapter 6 are associated. They represent two different experiments of the same academic system *ANACALYPSE* with different parameters affecting each experiment. The experimental results pertaining to Chapter 5 are presented in Appendix A – Data Table A and the experimental results pertaining to Chapter 6 are presented in Appendix B – Data Table B.

Furthermore, Chapter 5 describes the research background, the methodology and metrics and the relevant experimentation. In order for the reader to fully understand the information in Chapter 5 some knowledge of the topics in Chapter 6 is recommended. Hence in the experimentation section of Chapter 5 the reader is kindly referred to Chapter 6.

Furthermore, Chapter 6 describes the research environment, some information processing considerations, information analysis and machine learning techniques, similarity heuristics, the architectonic model and the statistical analysis of the relevant experimental data.

5.2 Introduction

Data storage technology and emphatically data capacity proliferates at an escalated rate of growth far surpassing Moore's law which estimates the growth rate for semiconductor manufacturing technology's transistor capacity (Gelsinger, Gargini,

Parker and Yu, 1989). As the amount of polymorphous data continues to increase there is a need for improved data information retrieval effectiveness and efficiency.

In this study syntheses of bidirectional fuzzy logic and information retrieval methods are developed, analyzed and evaluated by subject matter experts in order to elucidate the bases for improved effectiveness and efficiency of data information retrieval. Information search, retrieval and discovery are essential functionalities emphatically when very large scale databases are concerned (Ganesan, Garcia-Molina and Widom, 2003).

Distributing scientific information amongst researchers, initially from text oriented data sources, was the primary aim of the internet which was incarnated in the form of a collection of interconnected computer networks.

The evolution of information technology was the reason for a great number of the original technical problems to disappear. Initial limitations which dictated a monopoly on text restricted data sources were removed, opening the way for a variety of polymorphous data sources (Petratos, 2003).

However, although polymorphous data sources provided an unprecedented wealth of creation and freedom of expression they introduced additional problems.

Problems such as processing and storing increasingly heterogeneous data and adding to the complexity of existing information retrieval systems, problems which are only amplified as the number of interconnected computers is continuously increasing.

All machines which are added to the internet require an internet protocol address. Currently the length of existing internet protocol addresses is 32 bits, which allows for 2^{32} unique addresses or exactly 4,294,967,296 distinctly addressable machines.

The next generation internet addressing protocol, version six, will increase substantially the number of possible addresses, allowing for 128 bits address length or 2^{128} unique addressable machines, which is exactly 2^{96} times greater than the current possible number of unique addresses (Lawrence & Giles, 1999; 1998).

The synthesis of all these conditions forms a new technological ecosystem where the volume of data transfers is gigantic and distributed very large scale databases are endemic.

5.3 Research background

Contemporary research areas of information processing systems, emphatically system design as well as system effectiveness and efficiency with large scale data bases have attracted increased interest from scholars. Modern computing equipment may in fact be capable of alleviating and solving to a certain degree the information overload problem.

For instance, with the introduction of the short wavelength 405nm blue violet laser the capacity of optical data storage technology increases whilst the corresponding required capacity of electromagnetic disks analogously proliferates maintaining a

delicate balance between the two technologies whilst advancing the frontiers of information processing.

However, this creates the need for improved tools for analyzing, organizing, retrieving and locating specific information in response to search requests from a given user population in an expedient manner with little cost in time and effort.

This is the aim of this research for which syntheses of bidirectional fuzzy logic with information retrieval methods are developed in order to provide a novel bidirectional fuzzy information retrieval theory which may prove valuable to information retrieval.

Furthermore, due to a variety of factors, notably the emphatic interest of academics with semantic text content rather than non textual information, contemporary information retrieval systems still locate information relying upon the fundamental basis of a keyword formed query synthesized with string and pattern matching techniques (Lawrence, Giles & Bollacker, 1999).

Sceptical critics believe that there is nothing of merit to be discovered on the web. This thesis is particularly endemic amongst academic researchers. The antithesis is expressed by enthusiastic technocrats who hypothesize that every information requirement can be met simply by typing a few keywords into a web search engine and hoping that something relevant is returned.

The true gnomon has to be somewhere in between the thesis and antithesis expressed by the aforementioned dyad of opposite extremes. The principal pair of user activities

on the internet are communication and research (Cole, 2003; Lawrence & Giles, 1999; Kehoe & Pitkow, 1996).

Users can conduct research on a wide variety of topics. For instance, one topic of interest is to find out the newest developments about the war in the Middle East. Another topic of interest is to find out the locations and preferably instances of the most current scientific journal research papers or to find out about the most recent laboratory results of a new pharmaceutical medicine. A further topic of interest is to find out the most modern developments in image and face recognition algorithms or to find out details about a preferred computing product, etc.

Due to the vast amount of heterogeneous information on the web, which is characterized by an inherited lack of organization, users perform their research with the aid of web search engines. A series of internet surveys have been conducted in two research centres over a period of years with tens of thousands of subjects to historically track demographics and usage patterns on the internet (Cole, 2003; Kehoe & Pitkow, 1996).

The results of the surveys revealed that 84.8% of users utilize search engines to locate required information; hence it is not surprising that almost always search engines appear amongst the top accessed sites with the highest number of visitors (Cole, 2003; Kehoe & Pitkow, 1996).

The strategy of searching by keyword is characterized by numerous inelasticities. For instance, if the document sought is authored with a subtle variation in terminology

compared to the given keyword formed query then as a result the simple keyword search will fail to retrieve the document of interest.

Naturally, in order to accurately articulate and convey an information need, intelligent communication is required between the human and the machine. However, intelligent communication demands more interaction than a simple keyword formed query.

Even if the keyword formed query is successful a human must still read and discover the needed information somewhere in the myriads of documents matching the issued keyword formed query which is one of the aims of this study.

5.4 Methodology and metrics

Current research areas involving information processing systems can be broadly categorized into two categories, notably, information retrieval and knowledge discovery.

The former focuses on locating the relevant information entities among a large collection of potential contenders. The latter focuses on discovering new information from data mining using synthesis of events, detection of patterns or correlations among data in order to discover novel information which was not obvious or explicit previously.

These two broad research categories are endemic among a variety of information processing themes including information retrieval and extraction, text and data mining, topic detection and tracking.

Contemporary information retrieval systems utilize various scoring methodologies based upon statistical data of various parameters including term occurrences, encapsulated context, syntax, proximity to exoteric synonymous lexicon from a thesaurus, esoteric and exoteric link structure analysis (Brin & Page, 1998), etc.

All the aforementioned parameters are critical for different methodologies in order to compute the hierarchy of the search results. These document scoring techniques have two aims, firstly to ascertain the significance of each document rank within the specified query-resulting document-collection and secondly to reflect the credibility and importance of each document independently utilizing metrics such as number of citations, quality of reference sources, number of readers (Brin & Page, 1998), etc.

Furthermore, the effectiveness of information retrieval systems is evaluated based on the quality of results returned in response to a query. The most popular pair of metrics, which reflect the effectiveness of information retrieval systems notably are precision and recall.

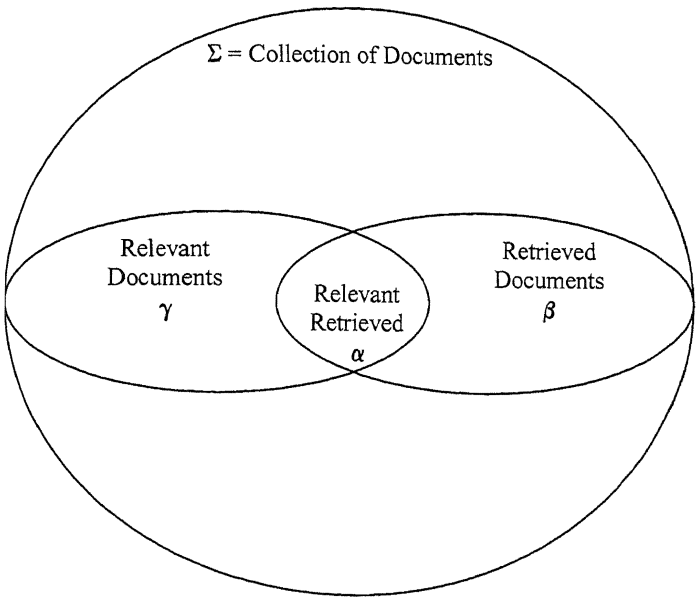

Figure 5.1. Illustration of effectiveness metrics for information retrieval systems.

In Figure 5.1 the set graph displays the parameters involved in the computation of these retrieval effectiveness metrics, notably precision and recall. Precision is a metric of the quality of the retrieved set of documents. Precision is equal to the fraction of the relevant documents retrieved from the entire set of retrieved documents.

$$P = \frac{\alpha}{\beta} \tag{5.1}$$

However, precision does not take into account the entire set of relevant documents which is a very important parameter for the effectiveness of an information retrieval system. On the other hand recall does take into account the entire set of relevant documents and is elucidated as follows. Recall is equal to the fraction of the relevant documents retrieved from the entire set of relevant documents existing in the whole collection.

$$R = \frac{\alpha}{\gamma} \tag{5.2}$$

For instance, if the set of retrieved documents is equal to the set of relevant documents retrieved then $\alpha = \beta$ and thus $P = 1$ hence this indicates information retrieval of absolute exactitude. However, if the entire set of relevant documents in the whole collection is equal to four times the set of relevant documents retrieved then $\gamma = 4 \times \alpha$ implying $R = 0.25$, hence this indicates information retrieval with a limited scope.

If the set of retrieved documents is equal to the set of relevant documents and thus equal to the set of relevant retrieved documents then this would be the perfect information retrieval system.

$$\beta = \gamma \Rightarrow \alpha = \gamma \Rightarrow \begin{cases} P = 1 \\ R = 1 \end{cases} \tag{5.3}$$

However, this is utopia considering that in reality information retrieval systems recover a large number of non-relevant documents. In reality precision of information retrieval fluctuates as recall changes and typically the former is reduced as the latter increases.

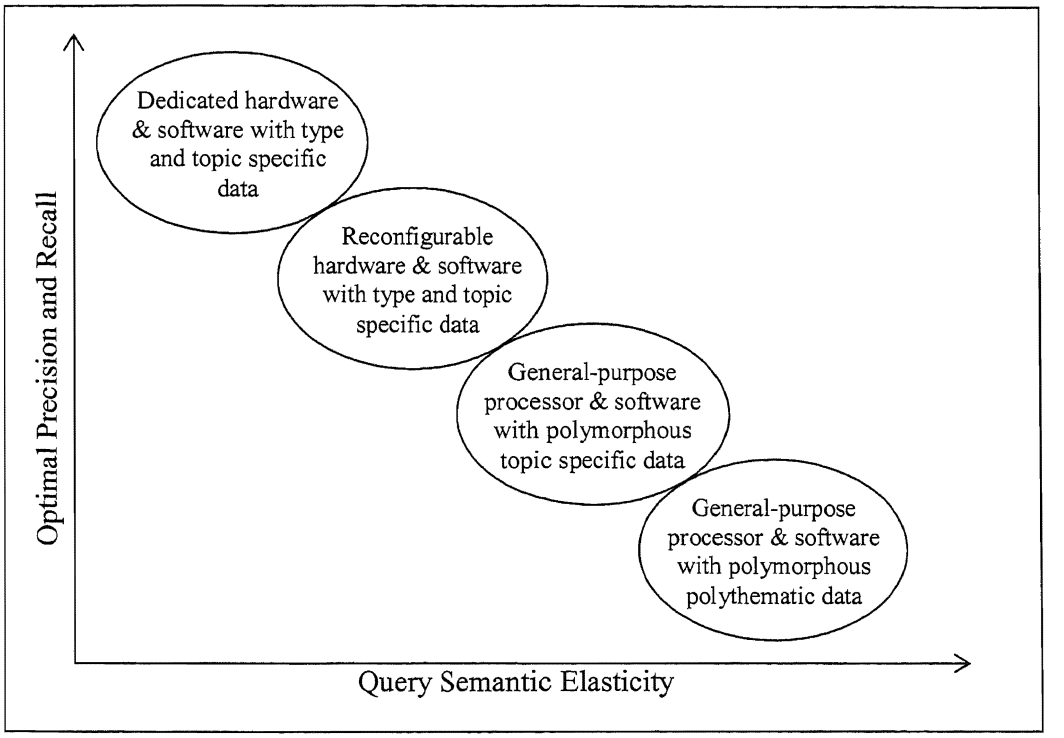

Figure 5.2. Graph of optimal precision, recall and query semantic elasticity correlations.

Furthermore, as shown in Figure 5.2, precision and recall also fluctuate according to the types of hardware, software and data of information retrieval systems. Early dedicated information retrieval machines with dedicated hardware and software utilized microcontrollers for specific purposes and functionality.

With the continuing evolution of nano-technology the scope, capacity and potential of nano-mechanics are steadily dilated (Witten, Moffat and Bell, 1999). Therefore, this integrated system on a chip design methodology is leading the metamorphosis of embedded microcontrollers with limited functionality into modern autonomous systems (Gelsinger, Gargini, Parker and Yu, 1989).

However, using a metaphoric phraseology, what good is the human body without the human soul and what good is the computer hardware without the necessary software? Based on the types of hardware, software and data of information retrieval systems there are certain correlations among the optimal precision and recall and the semantic elasticity of the query issued by the user.

Therefore, based on the previous correlations information retrieval systems can be broadly categorized into four taxonomies. Specifically, two categories are dedicated hardware and software with type and topic specific data, reconfigurable hardware and software with type and topic specific data.

Furthermore another two categories are general-purpose microprocessor and software with polymorphous topic specific data, general-purpose microprocessor and software with polymorphous polythematic data (Petratos, 2003). The correlations indicate that the more specific the types of hardware, software and data of the system the better the information retrieval effectiveness metrics precision and recall are whilst the worse is the semantic elasticity of the query.

On the other hand the more general the types of hardware, software and data of the system the better the semantic elasticity of the query, whilst the worse the information retrieval effectiveness metrics precision and recall are.

Analogously, contemporary information processing systems increasingly utilize polymorphous polythematic data (Petratos, 2003; Murray-Rust & Rzepa, 1999;

2002a; 2002b), which enrich the user experience and provide superior semantic elasticity for enquiries.

Furthermore, in modern general-purpose information retrieval systems, metrics precision and recall are equally influenced implicitly by the specific similarity coefficient which is used by the information retrieval system in order to determine the hierarchy of the entire document answer set.

In order to explicate a few of the well accepted similarity coefficients let a few propaedeutic parameters be briefly elucidated as they have already been presented earlier in Chapter 3.

A document is represented by a term vector consisting of a collection of terms. The process also known as *word stemming* or *lemmatization* extracts the roots of the terms (Porter, 1997; Salton & McGill, 1997; Salton 1989). Analogously, the aforementioned statements hold true as well for a query.

If a specific term is not present in the document or the query of interest, then the weight of that specific term is assigned a value of zero for the corresponding document or query vector respectively.

A collection of documents is represented by a matrix of term weights as displayed in formula 3.12 on page 66. Naturally, the same term normalization applies between document vectors as between document and query vectors.

In information retrieval, a well accepted geometrical computational model representing queries and documents by term sets in m-dimensional space is the vector space model (Salton, 1989; Salton, Wong and Yang, 1997).

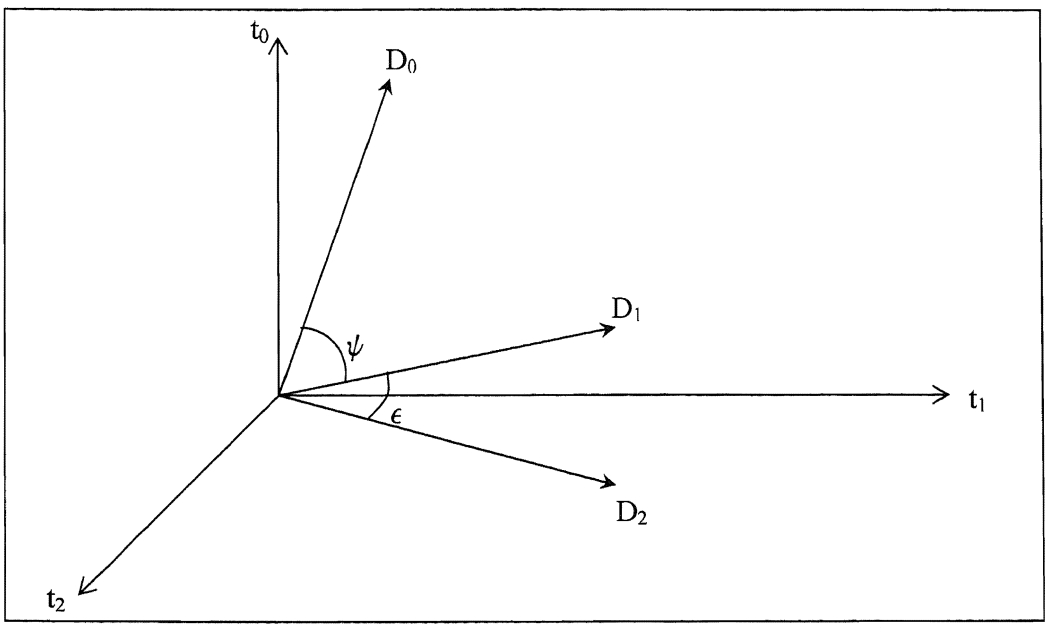

Figure 5.3. Graph of three dimensional vector space model illustrating document representation.

For instance, as shown in Figure 5.3, consider a triad of document vectors in a three term dimensional space and the corresponding angles between them ψ, ε then the document vectors can be visually portrayed as displayed in Figure 5.3.

The similarity between documents D_1 and D_2 is greater than the similarity between documents D_0 and D_1 due to the smaller angular distance between the vectors in the former rather than in the latter dyad.

Four of the most popular similarity coefficients are listed subsequently. The inner product similarity coefficient is the first (Wen, Nie & Zhang, 2002). For the inner product similarity coefficient formula the reader is referred to Chapter 4.

Furthermore, in the vector space model (Salton & McGill, 1997; Salton 1989) one of the well accepted similarity coefficients is the cosine between a dyad of vectors with a particular angular displacement. For the cosine similarity coefficient formula the reader is referred to Chapters 3 and 4.

Another well accepted similarity coefficient is the Dice metric (Ganesan, Garcia-Molina & Widom, 2003). For the Dice similarity coefficient formula the reader is referred to Chapter 4.

The fourth element of the tetrad is the Jaccard similarity coefficient (Ganesan, Garcia-Molina & Widom, 2003). For the Jaccard similarity coefficient formula the reader is referred to Chapter 4.

All these similarity coefficients provide a numerical quantification of the association of two documents or the arithmetical expression of the homology between a document and a query. Therefore, an information retrieval system can display retrieved documents in decreasing order of significance corresponding to their similarity with the query issued by the user.

Furthermore, the granularity of the hierarchical order of the documents in the entire answer set introduces an additional critical parameter which influences the

effectiveness of information retrieval systems. For instance, consider a system with two given different information retrieval strategies, notably *Method A* and *Method B* which yield document ranks, precision and recall as shown in Table 5.1 which follows.

Table 5.1. A dyad of different retrieval methods and their effects on metrics precision and recall.

Method A				*Method B*					
Rank	Doc #	Recall	Precision	Rank	Doc #	Recall	Precision	Recall Change	Precision Change
1	1	0.0000	0.0000	1	*2*	0.1428	1.0000	+14.28%	+ 100%
2	*2*	0.1428	0.5000	2	*5*	0.2857	1.0000	+14.29%	+50%
3	3	0.1428	0.3333	3	*8*	0.4285	1.0000	+28.57%	+66.67%
4	4	0.1428	0.2500	4	1	0.4285	0.7500	+28.57%	+50%
5	*5*	0.2857	0.4000	5	3	0.4285	0.6000	+14.28%	+20%
6	6	0.2857	0.3333	6	4	0.4285	0.5000	+14.28%	+16.67%
7	7	0.2857	0.2857	7	6	0.4285	0.4285	+14.28%	+14.28%
8	*8*	0.4285	0.3750	8	7	0.4285	0.3750	0%	0%
9	9	0.4285	0.3333	9	9	0.4285	0.3333	0%	0%
10	10	0.4285	0.3000	10	10	0.4285	0.3000	0%	0%
Mean	values:	0.2571	0.3110			0.3856	0.6415	+12.86	+32%

If the same query is issued for both *Method A* and *Method B* then the answer sets include ten documents for each method in the order displayed. A total of seven documents in the entire collection are relevant.

Only three relevant documents are retrieved for both methods out of the possible seven which is why the recall metric never reaches completeness failing to attain the maximum value possible.

However, the second method ranks the relevant documents in a more effective order than the first method which displays the relevant documents two five and eight in the identical rank position.

Conversely, the second method sequentially lists all the relevant documents at the top without any open space in between them, which consequently significantly improves precision, which reaches maximum for the first three relevant retrieved documents and analogously improves recall which reaches its highest value when the third relevant retrieved document is encountered and remains stable at the same value until the end of the answer set.

Another observation is that the lower the terminal relevant document is encountered in the first method the more data points improve their precision and recall in the second method.

For specific cases such as this when the ranked document answer set is considered in decreasing order of correlation between the documents and search requests alternative evaluation measures which take into consideration the rank of each document are in favour.

In addition to the typical precision and recall metrics, which depend on the number of retrieved documents, there are measures which are independent of the retrieved set size and are based on the system ranks such as formulas 5.4 and 5.5.

Such measures utilize system ranks of relevant documents compared with the ideal ranks of relevant documents by an ideal system where all relevant documents are encountered before any non-relevant documents appear.

Two metrics which exhibit the aforementioned characteristics, notably formulas 5.4 and 5.5 are based upon the information retrieval system document ranks and are expressed as follows (Salton & Lesk, 1997).

$$R_{norm} = \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n \cdot (N - n)} \tag{5.4}$$

$$P_{norm} = 1 - \frac{\sum_{i=1}^{n} \log r_i - \sum_{i=1}^{n} \log i}{\frac{\log N!}{(N-n)!\,n!}} \tag{5.5}$$

In formulae (5.4; 5.5) n is the number of the relevant documents, N is the number of the documents in the entire collection and r_i is the rank of the i-th relevant document as listed by the information retrieval system in decreasing hierarchy of the document correlation with the search request expressed by the user.

Naturally, the user preference provides a more accurate elucidation of the user information need rather than the more abstract idea of relevance. According to user preference there is a dyad of ranking strategies which exhibit a strong influence on the design of information retrieval systems, notably perfect and acceptable ranking (Wong & Yao, 1990).

The former ensures that the preferred documents are always ranked ahead of the less preferred and undesired documents. The latter only ensures that the less preferred documents are not ranked ahead of the preferred.

For instance, given a set of document vectors there is a corresponding preferred ranking generated by the user information need which dictates that the user prefers document five to document eight.

Furthermore, if the user is also a subject matter expert in the specific topic of the information search request then the corresponding preferred ranking generated is guaranteed to be an exhaustive unambiguous relevance ordering.

5.5 Experimentation

At this point it is noteworthy to indicate that as the architecture of *ANACALYPSE,* the academic experimental information retrieval system, is fully described in Chapter 6 the reader is referred to kindly see Chapter 6 in order to fully understand the experimental procedures described herein.

This novel bidirectional fuzzy logic theory is incarnated in an experimental information retrieval system, namely *ANACALYPSE* which is used for a series of experiments in order to determine the system's retrieval effectiveness.

The present study summarizes the results obtained with the *ANACALYPSE* information retrieval system based on the processing of *12,800* documents and presents evaluation output in comparison to a commercial information retrieval system, notably Google.

ANACALYPSE is an experimental information retrieval system encompassing bidirectional fuzzy expert relevance feedback, document representation in the vector

space model, and allowing for assessment by subject matter experts in order to determine retrieval effectiveness of the system.

Table 5.2. The heterogeneous user population selected for the experiments.

User ID	Male	Age: x>30	Native English
E1	1	0	1
E2	1	0	0
E3	1	1	0
E4	1	1	1
E5	0	1	1
E6	0	0	1
E7	0	0	0
E8	0	1	0

The Table 5.2 displays the user population selected for the information retrieval experiments. The user population is uniformly distributed among males, females, older and younger than thirty years of age, non-native and native English users.

The propaedeutic training set consists of a document collection with terms weighted according to expert bidirectional fuzzy relevance feedback given about the documents which are combined into a single vector called a *Metagram*, which is used for vector comparison with the unseen document vectors.

The comparison of *Metagram* with an unseen document vector is accomplished through vector analysis and similarity computation of the cosine between this dyad of vectors. All this information is stored and used to generate a knowledge base for further information retrieval sessions.

The term weighting approaches in *ANACALYPSE* incorporate a vector length normalization factor. The *Metagram* is a matrix vector structure, which is created by a combination of all the $n+1$ documents and their corresponding σ_i expert bidirectional fuzzy relevance feedback.

Numerous strategies have been reported regarding the elicitation of relevance information and the representation of meaning (Freeman, 2000; Kobayashi, Chang, & Sugeno, 2002). Herein a novel approach is developed whereby expert bidirectional fuzzy relevance feedback is elucidated on the basis of the cyclical thesis-antithesis as follows.

The thesis represents a position and the anti-thesis an opposite of the thesis. All the rest of the results are somewhere in between these two extreme locations in various angular displacements, with the majority being neutral occupying the area around the middle position.

At this point, it is noteworthy to remark that in the cases regarding a dyadic antithesis, the north and south poles are not involved. Also, note that while the antithesis location carries a -1 relevance feedback weight, the thesis location carries a 1 and, as the angle of the unseen document vectors decreases towards either extreme location, the similarity between the unseen document and the seen document represented by the either extreme locations increases.

Furthermore, according to Salton's cosine measure, one can easily see that although there is clearly an antithesis between the –x and east x positions, the y and –y positions are left unaffected since they have the same cosine value.

To illustrate this concept, let us consider an example. If there are numerous coordinates of the various communications range positions corresponding to the orbital trajectory of a geo-dynamic communications satellite then as the satellite moves around the earth different parts of the planet will have the ability to communicate with the satellite.

Hence if a user would like to find out exactly at which coordinates the satellite is active at any given time a geographical information retrieval system which would best match the active satellite position with the search request may prove valuable.

The experimentation occurred in a controlled environment with eight subject matter experts from the University of Luton to assess and evaluate the effectiveness of various possible search and analysis strategies.

The overall aim is to propose an effective methodology for automatic information retrieval systems, with semantically relevant output compared to the expert's "golden standard" (i.e. an expert's evaluation of relevance).

The experiments have two objectives:

1. To automatically compute the relevance of a large number of unseen documents based on expert relevance feedback on a small number of evaluated documents.
2. Moreover, to evaluate the experimental system *ANACALYPSE* and Google, the commercial information retrieval system, compared to the expert's relevance standard.

The expert-supervised training of the bidirectional fuzzy information retrieval system can be conducted in order to eventually achieve unsupervised system document relevance classification. This is based on an initial small training sample and a relevance function which is formed through a combination of the supervised training, the document similarity measure, and the term weighting function.

In order to measure the convergence of the experimental and the commercial information retrieval systems with respect to the experts' relevance standard, the Spearman's correlation coefficient is used which is defined as follows.

$$S_r = 1 - \left(\frac{6 \cdot \sum d^2}{n \cdot \left(n^2 - 1\right)} \right) \tag{5.6}$$

The procedure is based on a set of queries, which are selected in the research area of expertise. Research areas of expertise include artificial intelligence, bioinformatics, software engineering, information retrieval, decipherment, etc. The experts are researchers at the University of Luton. This approach is selected in order for the relevance training and Metagram generation to be accurate. The closer Spearman's correlation coefficient is to +1 the more linear is the relationship of the two ranks compared.

For each query the experimental system retrieved and processed 100 documents. Hence for the sum of 128 queries a grand total of *12,800* documents were retrieved and processed by *ANACALYPSE*. For this experiment the average Spearman's correlation of *ANACALYPSE* to the experts' relevance judgements was 0.45811. On the other hand the average Spearman's correlation of Google to the experts' relevance judgements was 0.190144.

According to these experimental results *ANACALYPSE,* the academic experimental information retrieval system is closer to the expert's relevance ranks with an average 12% positive change over Google, the commercial information retrieval system.

The data is shown analytically in Data Table A in Appendix A where a synopsis of the syntheses BAN-ABB-BBB from the experimental methodology is presented according to retrieval data for *12,800* documents processed.

Herein a statistical analysis of this data is presented. In the following tables and graphs the variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation.

In this phase *ANACALYPSE* uses the cosine similarity function. Naturally the higher positive the *AMINUSG* variable is, the greater is the distance between the academic and the commercial systems, whilst the closer are the academic system and the experts.

Table 5.3. Case processing summary of *AMINUSG* and *GENDER* cases.

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	GENDER	N	Percent	N	Percent	N	Percent
AMINUSG	0	64	100.0%	0	.0%	64	100.0%
	1	64	100.0%	0	.0%	64	100.0%

In this Table 5.3 a synopsis of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation.

As the reader would expect the higher positive the *AMINUSG* variable is, the greater is the difference between the academic and the commercial systems and the closer are the academic system and the experts' ranking. On the other hand the lower the *AMINUSG* variable is, the smaller the difference between the academic and the commercial systems.

The total number of cases presented herein consists of all valid cases and in-valid or missing cases. In this instance, all invalid or missing cases are zero. The purpose of this statistical processing is to study the effect of the gender of the user population on the correlation results.

Table 5.4. Descriptive statistics summary of *AMINUSG* and *GENDER* cases.

Descriptives

	GENDER			Statistic	Std. Error
AMINUSG	0	Mean		.209253	7.50E-02
		95% Confidence Interval for Mean	Lower Bound	5.94E-02	
			Upper Bound	.359100	
		5% Trimmed Mean		.187976	
		Median		9.09E-02	
		Variance		.360	
		Std. Deviation		.599885	
		Minimum		-.6908	
		Maximum		1.5393	
		Range		2.2301	
		Interquartile Range		1.054475	
		Skewness		.571	.299
		Kurtosis		-.733	.590
	1	Mean		.326680	5.36E-02
		95% Confidence Interval for Mean	Lower Bound	.219596	
			Upper Bound	.433763	
		5% Trimmed Mean		.323253	
		Median		.218200	
		Variance		.184	
		Std. Deviation		.428689	
		Minimum		-.4727	
		Maximum		1.1393	
		Range		1.6120	
		Interquartile Range		.733400	
		Skewness		.150	.299
		Kurtosis		-1.196	.590

In this Table 5.4 a synopsis of the descriptive statistics of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this statistical processing is to study the effect of the gender of the user population on the correlation results.

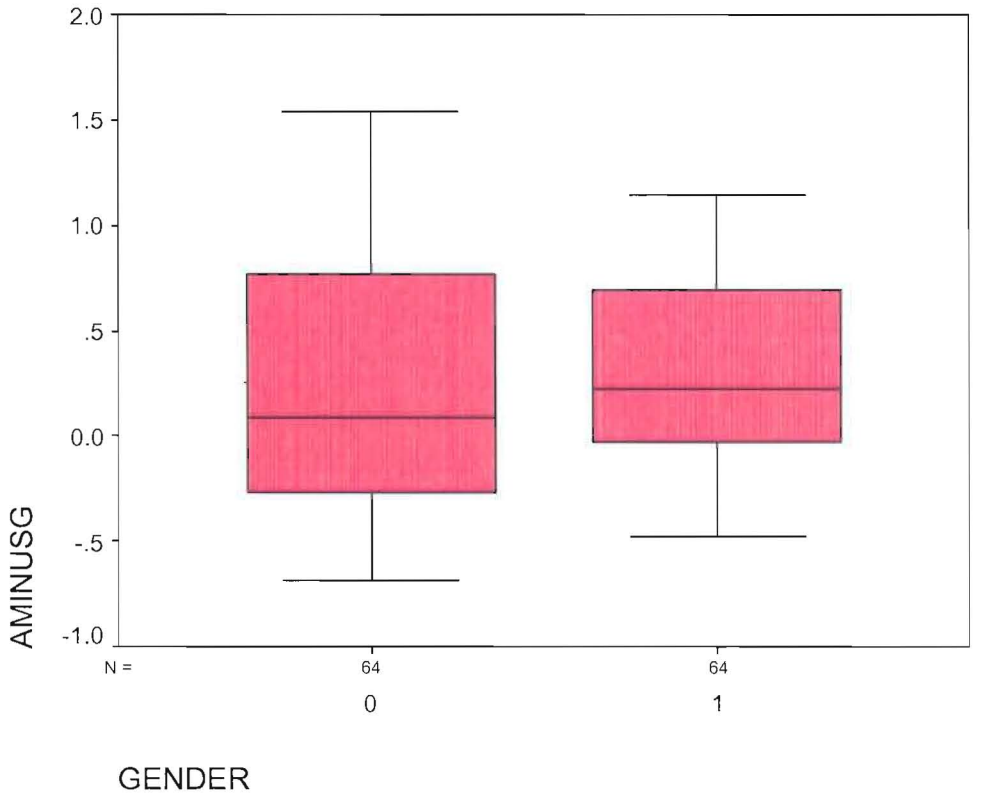

Figure 5.4. Box plot of *AMINUSG* and *GENDER* cases, zero indicates females, whilst one indicates males.

In this Figure 5.4 a box plot of the descriptive statistics of the *AMINUSG* and *GENDER* cases processed is presented. The y axis in this case represents the mean correlation *AMINUSG*.

The *GENDER* variable is zero for females and one for males. The descriptive statistics for females exhibit a wider distribution than for the males.

However, the mean value for females is slightly lower than for the males. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation

minus the *GOOGLE* correlation. The purpose of this statistical processing is to study the effect of the gender of the user population on the correlation results.

Table 5.5. Case processing summary of *AMINUSG* and *AGE* cases.

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	AGE	N	Percent	N	Percent	N	Percent
AMINUSG	0	64	100.0%	0	.0%	64	100.0%
	1	64	100.0%	0	.0%	64	100.0%

In this Table 5.5 a synopsis of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation.

As the reader would expect the higher positive the *AMINUSG* variable is, the greater is the difference between the academic and the commercial systems and the closer are the academic system and the experts' ranking. On the other hand the lower the *AMINUSG* variable is, the smaller the difference between the academic and the commercial systems.

The total number of cases presented herein consists of all valid cases and in-valid or missing cases. In this instance, all invalid or missing cases are zero. The purpose of this statistical processing is to study the effect of the age of the user population on the correlation results.

Table 5.6. Descriptive statistics summary of *AMINUSG* and *AGE* cases.

Descriptives

	AGE			Statistic	Std. Error
AMINUSG	0	Mean		.492380	6.96E-02
		95% Confidence Interval for Mean	Lower Bound	.353209	
			Upper Bound	.631551	
		5% Trimmed Mean		.493051	
		Median		.624200	
		Variance		.310	
		Std. Deviation		.557146	
		Minimum		-.4727	
		Maximum		1.5393	
		Range		2.0120	
		Interquartile Range		.987725	
		Skewness		-.101	.299
		Kurtosis		-1.322	.590
	1	Mean		4.36E-02	4.64E-02
		95% Confidence Interval for Mean	Lower Bound	-4.9E-02	
			Upper Bound	.136316	
		5% Trimmed Mean		4.25E-02	
		Median		6.67E-02	
		Variance		.138	
		Std. Deviation		.371358	
		Minimum		-.6908	
		Maximum		.7636	
		Range		1.4544	
		Interquartile Range		.466700	
		Skewness		.011	.299
		Kurtosis		-.532	.590

In this Table 5.6 a synopsis of the descriptive statistics of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this statistical processing is to study the effect of the age of the user population on the correlation results.

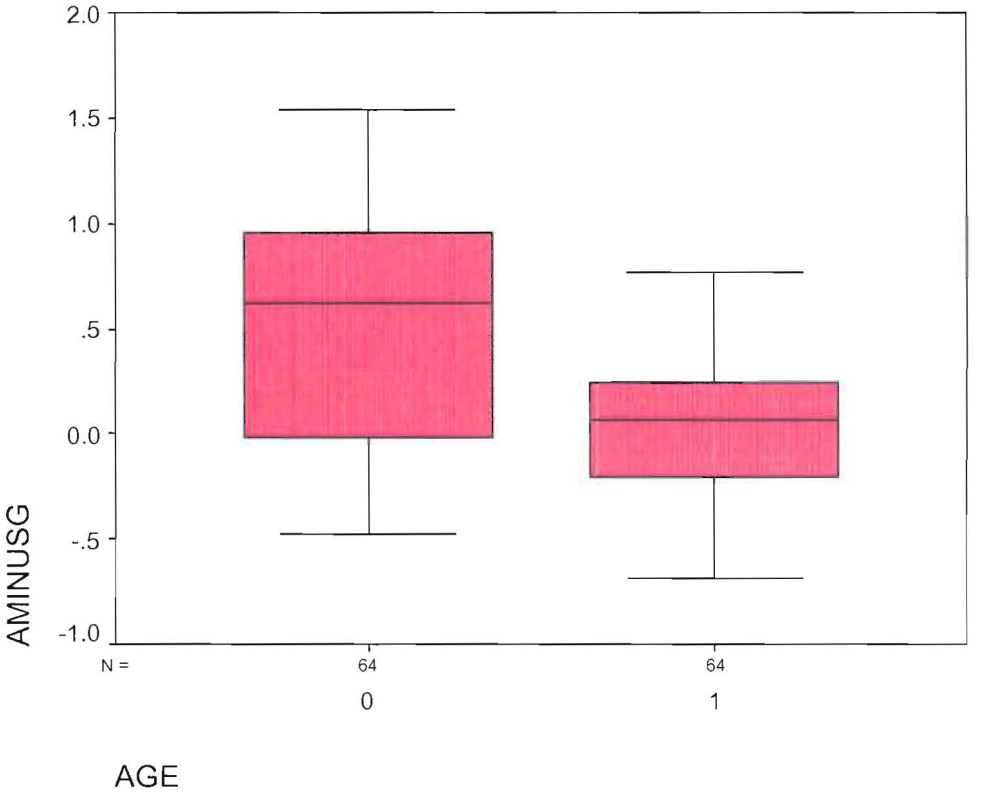

Figure 5.5. Box plot of *AMINUSG* and *AGE* cases, zero indicates ≤30 while one indicates > 30.

In this Figure 5.5 a box plot of the descriptive statistics of the *AMINUSG* and *AGE* cases processed is presented. The y axis in this case represents the mean correlation *AMINUSG*.

The *AGE* variable is zero for users of age less or equal to thirty and one for users of age greater than thirty years. The descriptive statistics for younger subjects exhibit a wider distribution than for the older subjects.

Furthermore, the mean value for younger subjects is slightly higher than for the older subjects. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this statistical processing is to study the effect of the age of the user population on the correlation results.

Table 5.7. Case processing summary of *AMINUSG* and *Native English* cases.

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Native English	N	Percent	N	Percent	N	Percent
AMINUSG	0	64	100.0%	0	.0%	64	100.0%
	1	64	100.0%	0	.0%	64	100.0%

In this Table 5.7 a synopsis of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation.

Naturally, the higher positive the *AMINUSG* variable is, the greater is the difference between the academic and the commercial systems and the closer are the academic system and the experts' ranking. On the other hand the lower the *AMINUSG* variable is, the smaller the difference between the academic and the commercial systems.

The total number of cases presented herein consists of all valid cases and in-valid or missing cases. In this instance, all invalid or missing cases are zero. The purpose of this statistical processing is to study the effect of the native tongue of the user population on the correlation results.

Table 5.8. Descriptive statistics summary of *AMINUSG* and *Native English* cases.

Descriptives

	Native English			Statistic	Std. Error
AMINUSG	0	Mean		.262673	6.19E-02
		95% Confidence Interval for Mean	Lower Bound	.139061	
			Upper Bound	.386286	
		5% Trimmed Mean		.265052	
		Median		.181800	
		Variance		.245	
		Std. Deviation		.494861	
		Minimum		-.6908	
		Maximum		1.1393	
		Range		1.8301	
		Interquartile Range		.821225	
		Skewness		.067	.299
		Kurtosis		-1.000	.590
	1	Mean		.273259	6.91E-02
		95% Confidence Interval for Mean	Lower Bound	.135150	
			Upper Bound	.411369	
		5% Trimmed Mean		.252739	
		Median		.127200	
		Variance		.306	
		Std. Deviation		.552898	
		Minimum		-.5091	
		Maximum		1.5393	
		Range		2.0484	
		Interquartile Range		.821250	
		Skewness		.555	.299
		Kurtosis		-.662	.590

In this Table 5.8 a synopsis of the descriptive statistics of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this statistical processing is to study the effect of the native tongue of the user population on the correlation results.

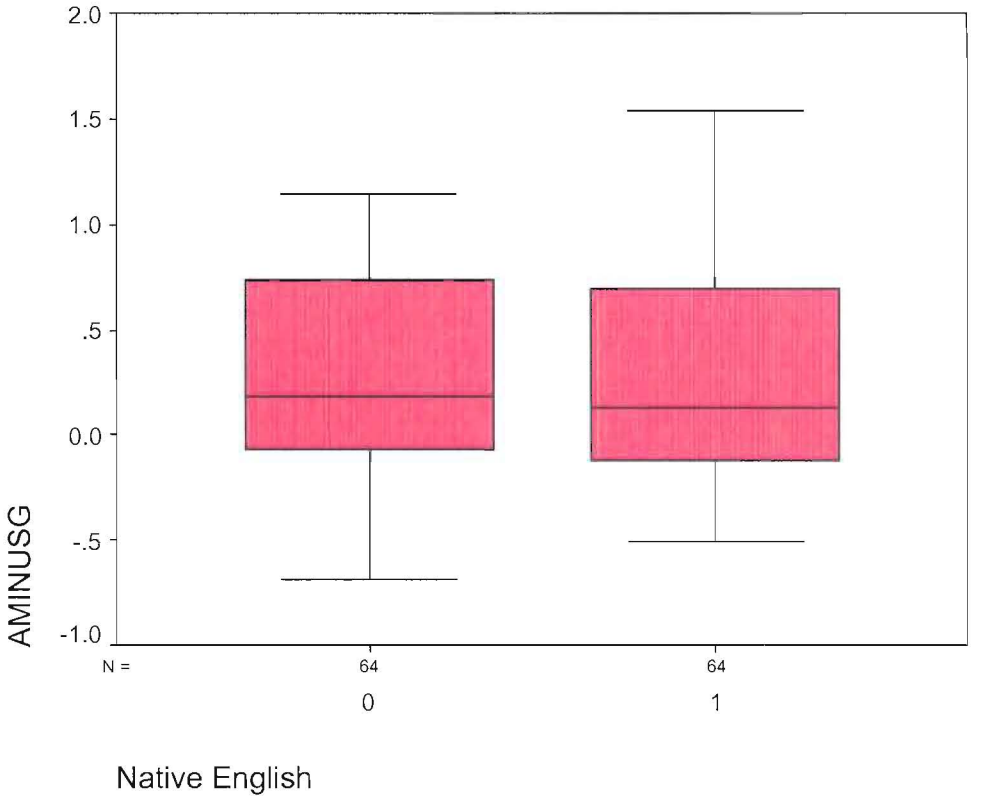

Figure 5.6. Box plot of *AMINUSG* and *Native English* cases, zero is no while one is yes.

In this Figure 5.6 a box plot of the descriptive statistics of the *AMINUSG* and *Native English* cases processed is presented. The y axis in this case represents the mean correlation *AMINUSG*.

The *Native English* variable is zero for users of native tongue other than English and one for users of English native tongue. The descriptive statistics for native English speakers exhibit the same distribution with the speakers of native tongue other than English.

Furthermore, the mean value for subjects of native tongue other than English is slightly higher than for the native English speakers. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this statistical processing is to study the effect of the native tongue of the user population on the correlation results.

Table 5.9. Case processing summing up of *AMINUSG* and stop words *REMOVAL* cases.

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	REMOVAL	N	Percent	N	Percent	N	Percent
AMINUSG	0	64	100.0%	0	.0%	64	100.0%
	1	64	100.0%	0	.0%	64	100.0%

In this Table 5.9 a synopsis of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation.

Unsurprisingly, the higher positive the *AMINUSG* variable is, the greater is the difference between the academic and the commercial systems and the closer are the academic system and the experts' ranking. On the other hand the lower the *AMINUSG* variable is, the smaller the difference between the academic and the commercial systems.

The total number of cases presented herein consists of all valid cases and in-valid or missing cases. In this instance, all invalid or missing cases are zero. The purpose of

this statistical processing is to study the effect of the stop-words removal process on the correlation results.

Table 5.10. Descriptive statistics summary of *AMINUSG* and stop words *REMOVAL* cases.

Descriptives

	REMOVAL			Statistic	Std. Error
AMINUSG	0	Mean		.350738	6.68E-02
		95% Confidence Interval for Mean	Lower Bound	.217319	
			Upper Bound	.484156	
		5% Trimmed Mean		.338364	
		Median		.218100	
		Variance		.285	
		Std. Deviation		.534118	
		Minimum		-.5091	
		Maximum		1.5393	
		Range		2.0484	
		Interquartile Range		.836375	
		Skewness		.419	.299
		Kurtosis		-.819	.590
	1	Mean		.185195	6.27E-02
		95% Confidence Interval for Mean	Lower Bound	5.99E-02	
			Upper Bound	.310450	
		5% Trimmed Mean		.179091	
		Median		.103000	
		Variance		.251	
		Std. Deviation		.501437	
		Minimum		-.6908	
		Maximum		1.1999	
		Range		1.8907	
		Interquartile Range		.866725	
		Skewness		.242	.299
		Kurtosis		-.972	.590

In this Table 5.10 a synopsis of the descriptive statistics of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this

statistical processing is to study the effect of the stop-words removal process on the correlation results.

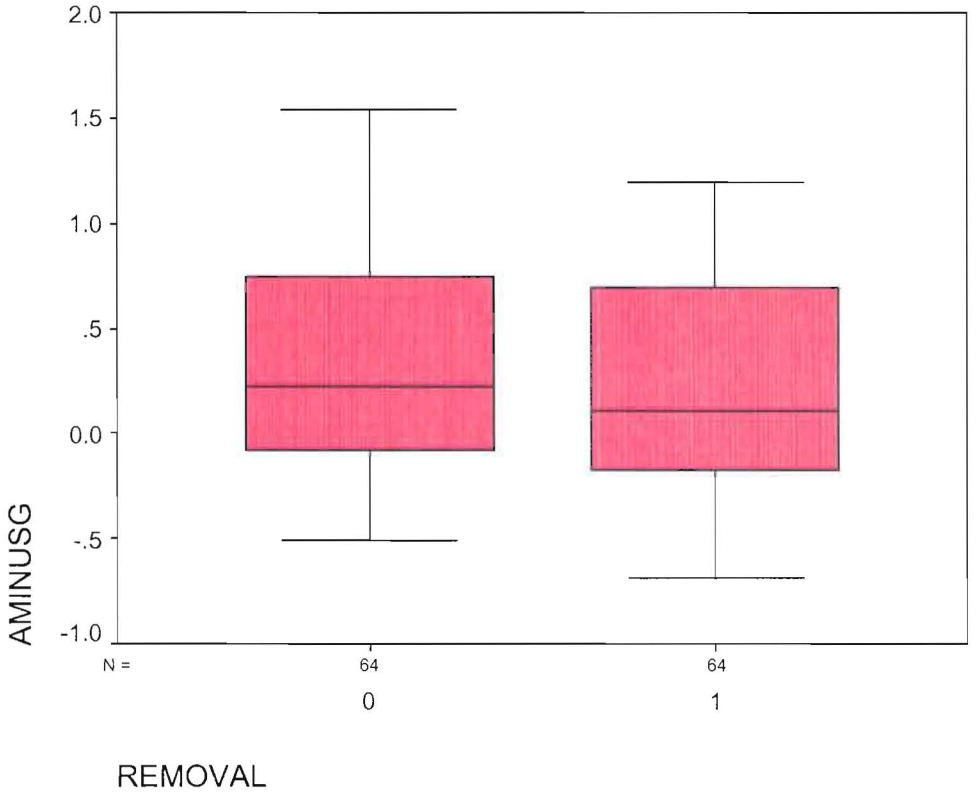

Figure 5.7. Box plot of *AMINUSG* and stop words *REMOVAL*, zero is no while one is yes.

In this Figure 5.7 a box plot of the descriptive statistics of the *AMINUSG* and *REMOVAL* cases processed is presented. The y axis in this case represents the mean correlation *AMINUSG*.

The *REMOVAL* variable is zero for documents which include the stop-words defined in Appendix C and one for documents which exclude these stop-words. The

descriptive statistics for documents including the stop-words exhibit the same distribution with the documents excluding the stop-words.

Furthermore, the mean value for documents including the stop-words is slightly higher than for documents excluding the stop-words. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this statistical processing is to study the effect of the stop-words removal process on the correlation results.

Table 5.11. Case processing summary of *AMINUSG* and *STEMMING* cases.

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	STEMMING	N	Percent	N	Percent	N	Percent
AMINUSG	0	64	100.0%	0	.0%	64	100.0%
	1	64	100.0%	0	.0%	64	100.0%

In this Table 5.11 a synopsis of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation.

Unsurprisingly, the higher positive the *AMINUSG* variable is, the greater is the difference between the academic and the commercial systems and the closer are the academic system and the experts' ranking. On the other hand the lower the *AMINUSG* variable is, the smaller the difference between the academic and the commercial systems.

The total number of cases presented herein consists of all valid cases and in-valid or missing cases. In this instance, all invalid or missing cases are zero. The purpose of this statistical processing is to study the effect of the stemming process on the correlation results.

Table 5.12. Descriptive statistics summary of *AMINUSG* and *STEMMING* cases.

Descriptives

	STEMMING			Statistic	Std. Error
AMINUSG	0	Mean		.274791	6.58E-02
		95% Confidence Interval for Mean	Lower Bound	.143268	
			Upper Bound	.406314	
		5% Trimmed Mean		.264794	
		Median		.157600	
		Variance		.277	
		Std. Deviation		.526529	
		Minimum		-.6908	
		Maximum		1.5393	
		Range		2.2301	
		Interquartile Range		.872700	
		Skewness		.328	.299
		Kurtosis		-.806	.590
	1	Mean		.261142	6.53E-02
		95% Confidence Interval for Mean	Lower Bound	.130552	
			Upper Bound	.391732	
		5% Trimmed Mean		.248831	
		Median		.127250	
		Variance		.273	
		Std. Deviation		.522793	
		Minimum		-.6787	
		Maximum		1.5393	
		Range		2.2180	
		Interquartile Range		.794025	
		Skewness		.385	.299
		Kurtosis		-.695	.590

In this Table 5.12 a synopsis of the descriptive statistics of the cases processed is presented. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this

statistical processing is to study the effect of the stemming process on the correlation results.

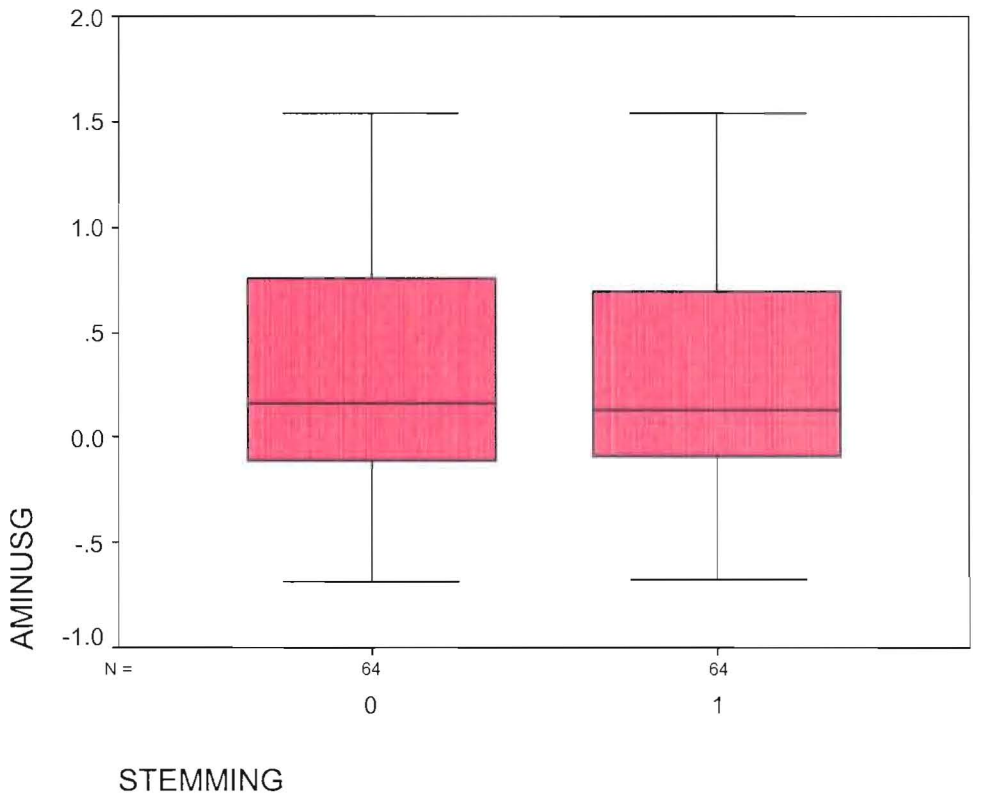

Figure 5.8. Box plot of *AMINUSG* and *STEMMING*, 0 is no while 1 is yes.

In this Figure 5.8 a box plot of the descriptive statistics of the *AMINUSG* and *STEMMING* cases processed is presented. The y axis in this case represents the mean correlation *AMINUSG*.

The *STEMMING* variable is zero for documents which exclude the stemming process defined in Appendix D and one for documents which include the stemming process. The descriptive statistics for documents excluding the stemming process exhibit slightly wider distribution than the documents including the stemming process.

However, the mean value for documents excluding the stemming process is slightly higher than for documents including the stemming process. The variable *AMINUSG* represents the difference given by the *ANACALYPSE* correlation minus the *GOOGLE* correlation. The purpose of this statistical processing is to study the effect of the stemming process on the correlation results.

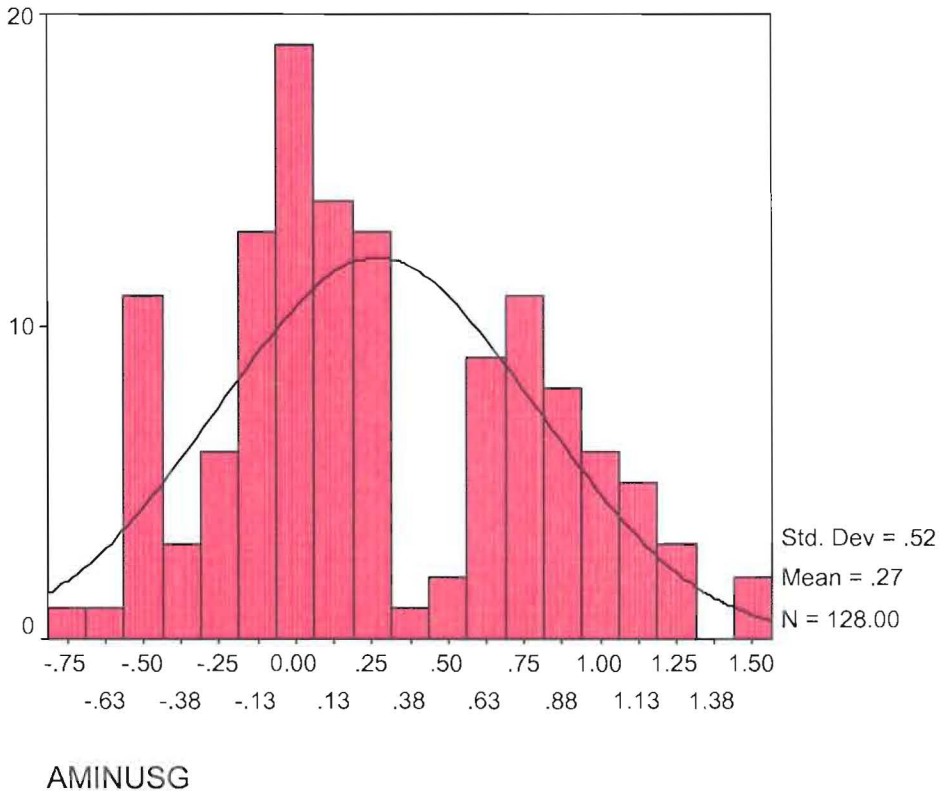

Figure 5.9. Histogram of *AMINUSG* showing distribution and normal curve.

This Figure 5.9 portrays the distribution of the values attained by the variable *AMINUSG* and the standard deviation, the mean and the total number of cases. In this specific instance the normal curve is approximately followed by the actual *AMINUSG* curve.

Table 5.13. T-test one sample *AMINUSG* statistics summary.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
AMINUSG	128	.267966	.522640	4.62E-02

In this Table 5.13 a synopsis of the one sample statistics of the variable *AMINUSG* is presented. The one sample statistics include the total number of cases, the mean, the standard deviation and the standard error mean values.

Table 5.14. One sample T-test *AMINUSG* statistics synopsis.

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
AMINUSG	5.801	127	.000	.267966	.176554	.359378

In this Table 5.14 a synopsis of the one sample T-test statistics of the variable *AMINUSG* is presented. The one sample T-test statistics include the total number of cases, the mean difference, and the lower and upper confidence interval of the difference values.

Table 5.15. Non parametric one sample K-S test *AMINUSG* statistics.

One-Sample Kolmogorov-Smirnov Test

		AMINUSG
N		128
Normal Parameters[a,b]	Mean	.267966
	Std. Deviation	.522640
Most Extreme Differences	Absolute	.121
	Positive	.121
	Negative	-.077
Kolmogorov-Smirnov Z		1.370
Asymp. Sig. (2-tailed)		.047

a. Test distribution is Normal.

b. Calculated from data.

In this Table 5.15 a synopsis of the non parametric one sample Kolmogorov-Smirnov test statistics of the variable *AMINUSG* is presented. The one sample Kolmogorov-Smirnov test statistics include the total number of cases, the mean, the standard deviation, and the most extreme positive and negative difference values.

Table 5.16. Non parametric signed ranks *AMINUSG* statistics.

Ranks

		N	Mean Rank	Sum of Ranks
GOOGLE - Anacalypse	Negative Ranks	84[a]	73.10	6140.00
	Positive Ranks	44[b]	48.09	2116.00
	Ties	0[c]		
	Total	128		

a. GOOGLE < Anacalypse

b. GOOGLE > Anacalypse

c. Anacalypse = GOOGLE

In this Table 5.16 a synopsis of the non parametric Wilcoxon signed ranks statistics of the variable *AMINUSG* is presented. The Wilcoxon signed ranks statistics include the

total number of cases, the mean rank, the positive and negative ranks and the sum of ranks.

Table 5.17. Non parametric Wilcoxon test *AMINUSG* statistics.

Test Statistics[b]

	GOOGLE - Anacalypse
Z	-4.785[a]
Asymp. Sig. (2-tailed)	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

In this Table 5.17 a synopsis of the non parametric Wilcoxon signed ranks test statistics of the variable *AMINUSG* is presented. From the above statistics it is easy to see that younger subjects seem to be more satisfied with the results of *ANACALYPSE.*

Furthermore, it appears that males are more satisfied with the results of the *ANACALYPSE* process than females. Stop-word removal, which is almost a universal norm in information retrieval, does not seem to be beneficial and can sometimes be harmful.

System stemming does not seem to be harmful or beneficial; it is simply neutral to the effectiveness of the information retrieval process. Furthermore the *AMINUSG* data histogram shows a slight deviation from the normal curve.

The t-test assumes that the data are normally distributed however the t-test is fairly robust to departures from normality as well. The one sample t-test procedure tests

whether the mean of a single variable differs from a specified constant. The one sample t-test results are significant.

Furthermore as there is a slight departure from the normal curve a non parametric one sample Kolmogorov Smirnov test should be applied. The Kolmogorov Smirnov test results again are significant.

Finally the non parametric Wilcoxon signed ranks test shows $P<0.0005$ a very highly significant difference.

While current commercial information retrieval systems provide quite adequate methods of general heuristic for information they can be limited in effectiveness and some times restrictive in query elasticity, conditions which represent additional effort in time and reading for the seeker.

In this study a synthesis of bidirectional fuzzy logic and information retrieval methods have been developed, analyzed and evaluated by subject matter experts. The novel methodology has been designed and implemented in an experimental information retrieval system.

According to the experimental results the academic information retrieval system is significantly closer to the expert's relevance judgements compared to Google, the commercial information retrieval system. Therefore the proposed methodology elucidates the bases for improved effectiveness and efficiency of data information retrieval.

CHAPTER 6 – Architecture of system and experimentation

6.1 Introduction

As far as the information science field is concerned, research endeavours in the general areas of information analysis, transformation, management, exploration, processing, and retrieval have recently attracted a surge of academic interest gaining a prominent research status emphatically due to the continuous increase of the availability of documents in digital form and consequently the elastic user access requirements stemming from this metamorphosis.

In addition, the transition from analogue to digital technology is increasingly influencing every aspect of modern life. Contemporary machines encompass progressively more subsystems composed with digital electronics and consequently the availability of digital information is experiencing a continuous growth.

Furthermore, digital information is not only increasingly expanding but is also increasingly diverse, covering a range of topics (Murray-Rust & Rzepa, 1999). The rising popularity of the internet and a variety of information services offering polymorphous data also result in ever-increasing digital information (Murray-Rust & Rzepa, 2002a; 2002b).

This considerable rate of growth which characterizes digital information is the origin of the well recognized information overload problem.

In addition this upsurge of digital information articles creates another challenge for information retrieval systems as it is inversely related to their effectiveness to locate the desired digital information articles from a very large collection of possible candidates (Ganesan, Garcia-Molina & Widom, 2003; Lawrence, Giles & Bollacker, 1999).

6.2 Research environment

In recognition of the problem of information overload, the research and development of systems to automatically analyze and explore information has become the focus of considerable interest and investment in both research and commercial fields (Cole, 2003; Lawrence & Giles, 1999; 1998; Kehoe & Pitkow, 1996).

This is the aim of information retrieval systems which allow the searcher to locate the most relevant, key information from a large collection of literature for a particular enquiry. However, when an information retrieval system either retrieves a collection of documents or recovers their possible locations in response to a searcher's request it is not unusual for a large fraction of the resulting set of documents to be non-relevant to the search request.

Therefore the information searcher is left with the burden of exhaustively reading in detail documents that may or may not be relevant, consequently spending a significant amount of time and effort fruitlessly. Furthermore, the strategy of searching by keyword is characterized by numerous inelasticities.

For instance, if the document sought is authored with a subtle variation in terminology compared to the given keyword-formed query then as a result the simple keyword search will fail to retrieve the document of interest. Naturally, in order to accurately articulate and convey an information need, intelligent communication is required between the human and the machine.

However, intelligent communication demands more interaction than a simple keyword-formed query. Even if the keyword-formed query is successful a human must still read and discover the needed information somewhere in the myriads of documents matching the issued keyword formed query.

These problems motivate this research, which seeks to investigate methods of improving the effectiveness of current information retrieval systems. This principal aim can be achieved by developing a number of methods and pursuing several subsidiary objectives.

One of the methods is introducing a novel bidirectional, symmetrical fuzzy logic theory which may prove valuable to information retrieval and exploration research and development (Petratos & Chen 2002; Petratos, Chen, Wang & Forsyth 2002).

Another method is synthesizing and applying bidirectional, symmetrical fuzzy logic to classic information retrieval theory. Furthermore, an experimental information system is to be designed and implemented, combining bidirectional fuzzy logic in order to automatically compute the relevance of a large number of unseen documents from expert relevance feedback on a small number of documents read.

The system is designed in such an approach that the evaluation of the effectiveness of the experimental information retrieval system is considered by allowing triadic comparisons to commercial information retrieval systems and in parallel to the expert's gold standard of judged relevance according to established statistical correlation methods.

6.3 Information processing considerations

Information retrieval research and development over the years have produced a variety of methods for locating information of interest utilizing robust text processing techniques.

A number of these methods take advantage of the text characteristics, automatically create categorized clusters of documents, search entire collections of documents and retrieve pertinent information based on incisive characterizations of their texts (Paice, 1990; Van Rijsbergen, 1979).

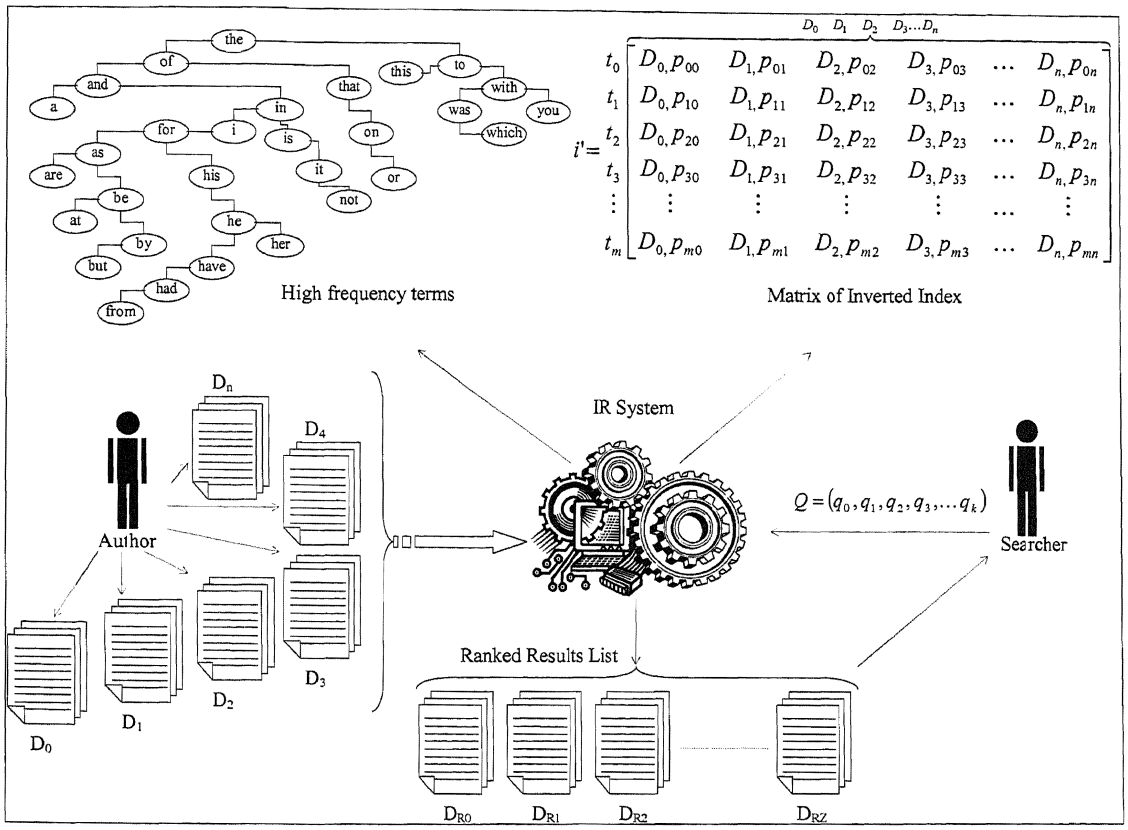

Figure 6.1. Graph of classical information retrieval systems research and development methodologies.

This Figure 6.1 portrays a diagram of classical models of information retrieval systems. Early research endeavours concerning the field of information retrieval had limited scope and were characterized by their simplicity of processing since during that epoch neither large corpora of electronic text, nor sophisticated natural language processing methods, nor powerful computers with large memory for fast indexing and searching existed.

These pioneering research endeavours investigated various methodologies including the not so unusually content revealing topical position of phrases in privileged locations such as the first paragraph or immediately following section headings or the

first and last sentences of each paragraph, the occurrences of lexical cues such as significant, hardly, impossible, which indicate topic related phraseology.

Naturally all these methodologies, although they can be very effective, they are still dependent upon the specific rhetorical style and type of authorship. For instance the strategy of analyzing the first paragraph can be very effective primarily in the genre of news articles.

Furthermore, during that epoch if a strategy failed to obtain the desired result due mainly to hardware limitations there were no automatic alternative techniques developed to illuminate optimal positions, relevant cues, relative distance of locations, etc.

Automated machine text reading and understanding techniques are likely to remain for a long period of time simple approximations of human reading and understanding. True text reading and understanding requires a higher level of sceptical analysis and interpretation of the text into a new synthesis at different levels of abstraction. The missing anthropologic analyzing capabilities were sought in artificial intelligence semantics-based methodologies.

For instance, one approach is to represent a series of high-level interpersonal interactions between protagonists called plot units into a scenario as an interconnected network and simply identify and capture the central idea of the action by selecting units according to the arithmetic count of interconnections from each plot unit to its

neighbours forming a chain of esoteric central plot units and eliminating the exoteric peripheral plot units (Lehnert, 1983).

However, plot units are very general interactions such as denied request, resign, success through adversity, and consequently they only provide the necessary means for rather abstract representation schemes. Other strategies instead of plot units employ frames or templates which contain a wider variety of objects and states of affairs (Rau & Jacobs, 1991).

Templates are also used as guiding blue-prints to perform automatic information identification and extraction i.e. from news texts in specific topics of interest such as terrorism. Information systems which utilize templates detect and extract through a variety of methods the prescribed types of sought information (Jacobs & Rau, 1990).

However a system with a static template produces consistently predictable output which is limited to a single topic according to the prescribed guide lines of the template and cannot serve as a general solution for all genres and topics.

Antithetically the more traditional, pure information retrieval approaches focus on the statistical data at the level of terms eschewing any further symbolic abstract semantic representations. This research approach presents some advantages as well as some interesting dilemmas (Rillof & Lehnert, 1994; Mauldin, 1991).

Although statistical term level techniques are axiomatically unambiguous and they have been well developed and applied in numerous practical information systems

applications there are certain cases where semantic concepts diverge from simple word sequence expressions.

For instance, if a semantic concept is expressed by different words such as cycle, bicycle, tricycle which all can refer to a vehicle or a geometric shape then this is called synonymy (Hudson, 1995).

Furthermore polysemy occurs if one word has several meanings such as life cycle, bicycle and vicious cycle. If a cycle query is issued then documents containing all the above will be returned. Another instance of polysemy is linked to syntheses of verbs such as take place, take a picture, take notice, take time and take medicine.

If a take query is issued then documents containing all the above will be returned. Finally phrases can have different meanings from the words they are consisted of, such as an alleged murderer is not a murderer until proven guilty, or Abraham has a Lincoln automobile, or the mother urgently told the child to duck away from the bullets.

In the first instance if a guilty murderer query is issued the paradigmatic documents will be returned. In the second instance if an Abraham Lincoln biography query is issued the paradigmatic documents will be returned. In the third instance if a duck hunting query is issued the paradigmatic documents will be returned.

Term dependencies, semantic phrases, synonymy, polysemy, are all concerns related to semantics. These concerns attracted the interest of researchers and numerous

approaches are developed relating these issues. One approach is to utilize an approximation to world knowledge consisting of pre-compiled, pre-existing online semantic knowledge sources such as word lists, thesauri and dictionaries.

For instance synonymy can be identified through the use of a thesaurus and a sense disambiguation algorithm can select the correct sense of a polysemous word (Yarowsky, 1992). Another approach is to identify phrase segments and use them as synonymous terms by employing a syntactic parser (Lewis, 1992). Another approach to address specifically term dependencies is to employ latent semantic indexing (Deerwester, Dumais, Landauer, Furnas and Harshman, 1990).

All these research endeavours have encouraged other researchers to continue and expand this work by synthesizing term level statistical techniques with epiphanic semantic processing in order to improve efficiency and effectiveness of automatic text categorization systems (Liddy, Paik & Yu, 1994).

All these approaches are noteworthy efforts to address the differences between term expressions and term meanings however machine understanding and learning still rudimentarily remains an approximation of the human ability to read and understand.

6.4 Information analysis and machine learning techniques

Information systems analyze information at a varied level of sophistication which can fluctuate widely. For instance, information systems can process lexical input from a document and produce a wide range of output such as a simple list of isolated key words which are reflective of the most important content of the document, or a

sequence of independent phrases which synthesized characterize in a coherent and integrated approach the most important content of the document.

The higher the level of sophistication of an information system the more complex the system processing becomes and the more computational effort and time is required. Early research endeavours in the areas of information retrieval, processing and extraction basically involved lexical and spatial topical analyses of the text utilizing statistical parameters such as frequency of key words, term proximity and location within the text (Luhn, 1958; Edmundson, 1969; Rush, Salvador & Zamora, 1971).

More recent research endeavours approach the problem with a slightly different strategy involving automated methods to synthesize these types of feature sets through classification techniques or alternatively research strategies are based upon traditional information retrieval indexing methods in order to incorporate knowledge of a text corpus (Brandow, Mitze & Rau, 1995).

Information retrieval, metamorphosis and analysis research as well as analogous information systems development methodologies can be broadly categorized in a triad of taxonomies. These three general categories are frequency centric, knowledge based and discourse centric research and systems development methodologies.

There are certain correlations between these categories of research and systems development methodologies and the progress of automated text processing and understanding. In the continuum formed by all the aforementioned parameters the

three general categories are analogous to increasing understanding of text as well as increasing system processing complexity.

The first category of information research and systems development methodologies is concerned with frequency centric systems. Frequency centric systems often involve heuristics and more traditional lexical statistical data and parameters in order to perform automatic metamorphosis, analysis and retrieval of information (Salton & McGill, 1997; Salton & Lesk, 1997; Salton, 1989; Porter, 1997).

In addition frequency centric systems are often used with free text data bases and rely on natural language for query interfaces hence measuring system effectiveness is rather subjective depending on how satisfying the answer set is for the searcher according to her request. Furthermore frequency centric systems are also employed to address readability related issues (Brandow, Mitze & Rau, 1995).

The second category of information research and systems development methodologies is concerned with knowledge based systems. In order to interpret the ideological structure of the text, knowledge based strategies often rely on a large number of rich domain knowledge sources.

Numerous experimental knowledge based information systems process texts of a specific domain in order to extract their corresponding ideological portrayals (Fum, Guida & Tasso 1985; Rau, Jacobs & Zernik, 1989).

These knowledge based information systems apply knowledge extracted from the domain in order to characterize specific implicit conceptual knowledge of a text often employing methodologies of automatic identification of relevant ideas highly correlated with a category of interest (Paice & Jones 1993; Riloff 1995). It is not unusual for knowledge based strategies to be explicitly domain specific with respect to the automatic extraction of conceptual representations of texts.

The third category of information research and systems development methodologies is concerned with discourse centric systems. Strategies of discourse centric systems are based in text cohesion and coherence theories which concern semantic lucidity. Discourse centric systems considerably fluctuate in their capacity variance of text understanding and the complexity as well as automation of system processing.

In contrast to frequency centric systems which do not take into consideration cohesion and coherence, discourse centric systems concentrate on linguistic text processing in order to identify the optimum cohesive phraseology or the optimum lexical phrase candidates for portraying the rhetorical structure of the text of interest. Discourse centric approaches rely on processing the text and analyzing discourse correlations in order to characterize the optimum semantic phraseology (Paice 1990; Johnson et al. 1993).

Although knowledge based strategies perform text analysis by carrying out discourse processing systematically, in a broad classification scheme discourse centric strategies place in the centre of attention the macrostructure and composition of the text. Alternatively a slightly different approach is to capture information about both

domain and structure of the text by taking advantage of differences and parallelisms of lexical and phrasal conceptual sources within a document (Mani & Bloedron, 1997).

Also, another approach is to represent the text of a source with respect to a manual method based on linguistic, domain and communicative descriptive information (Jones, 1995). Furthermore, there are certain correlations between the types of the documents being processed and the specific approach selected to perform the document analysis.

Depending on the actual synthesis of these parameters information processing systems may differ in several ways. For instance, some may include indicative key words reflecting topics, detailed information about the contents of documents, information according to the author's perspective, user specific query oriented inputs, analysis of genre specific documents i.e. solely news articles, non intrusive neutral analysis or pervasive evaluative analysis (Wen, Nie & Zhang, 2002; Wong & Yao, 1990).

A full understanding of the principal dimensions of variation and the corresponding reasoning types necessary to create each of them is still a matter of investigation which makes the study of metamorphosis, analysis and retrieval of information an exciting research area to investigate.

In the field of computational linguistics and emphatically in natural language processing there are numerous robust and established techniques such as corpus based

statistical lexical processing, automatically deriving sets of characteristic features of phraseology, automatically detecting and extracting information streams.

However, the majority of existing natural language processing techniques are more likely to operate at the word level and consequently miss concept level generalizations of semantic meanings.

These data parameters missed are provided by symbolic world knowledge which is embodied in online repositories of knowledge such as the lexical network and concept thesaurus WordNet, the categorized organizations of concept correlations in Google sets, and numerous other visual thesauruses and electronic dictionaries (Mihalcea & Moldovan 2001; Harabagiu & Moldovan, 1997; Beckwith & Miller, 1990).

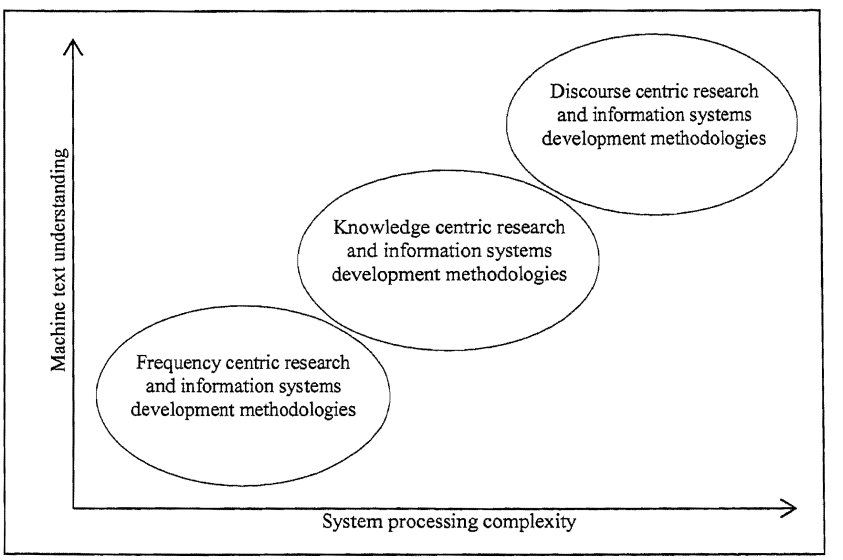

Figure 6.2. Graph of information systems research and development methodologies correlations.

This Figure 6.2 portrays a diagram of the correlations among machine text understanding and system processing complexity. Furthermore another interesting approach in semantic lexical categorization is to distinguish a topic hierarchy by analyzing word coincidences which form a distinct taxonomy by employing machine learning techniques from an initial training set in order to categorize subsequent documents based on the sets of words they contain.

From a categorization such as this a machine should be able to compute the subcategory for which the group of words are the best classifiers, and additional words would then be suggested by supplying additional classifiers for that category (Chakrabarti, Dom, Agrawal & Raghavan, 1998).

In general the machine learning paradigm encompasses a general inductive process which automatically creates an automatic text classifier for a category c_j by learning from a sample training set of documents manually classified by domain experts under c_j, the characteristics that a new unseen document ought to have in order to be classified under c_j.

Therefore, automatic text categorization in machine learning systems occurs as a supervised learning activity as the learning process is supervised by the knowledge gained from the training sample set of documents and their respective expert categorization. Thus automatic text categorization is elucidated as the assignment of a hard true or false Boolean value to each document-category dyad.

$$\{d_i, c_j\} \in D \times C \tag{6.1}$$

$$C = \{c_0, c_1, c_2, c_3 \ldots c_k\} \tag{6.2}$$

$$D = \{d_0, d_1, d_2, d_3 \ldots d_n\} \tag{6.3}$$

A decision to categorize document d_i under category c_j is indicated by an assigned value of T and the antithesis is indicated by an assigned value of F for each document-category dyad. Furthermore the machine learning paradigm for automatic text categorization is more formally elucidated as the approximation of the unknown terminus ad quem function.

$$\overset{\circ}{\Phi} : D \times C \rightarrow \{T, F\} \tag{6.4}$$

Hence (4) elucidates the optimal and ideal categorization of a collection of documents by means of another classifier function.

$$\Phi : D \times C \rightarrow \{T, F\} \tag{6.5}$$

Ideally if the classifier (6.5) gives the same results as the target function (6.4) then the machine learning system is of comparable effectiveness to human subject matter experts. However, in reality it is not unusual for a machine learning system to be designed and operate under the hypothesis that the approximation of the classifier (6.5) coincides as much as possible to the target function (6.4).

Naturally the selected training sample and the training strategy for the classifier are critical for the development and effectiveness of the machine learning system and it's esoteric classifier. Thus it is not unusual for the internal parameters of the classifiers to be tuned by iterative training sessions in order to yield the best effectiveness possible on the training data.

However this approach of relentless optimization introduces a phenomenon called over-fitting by which a classifier is tuned not only to the desired constitutive characteristics but also to the undesired contingent characteristics of the training data and by association of the classified categories.

Once a classifier is over fitted it exhibits good performance for re-categorization of the data it was trained on albeit it exhibits rather poor performance for categorization of unseen data.

Two strategies are employed in order to avoid over-fitting, the first is to use training paradigms analogous to the number of term dimensions and the second is to use term dimensionality reduction which effectively reduces the number of terms as well as the number of training paradigms (Fuhr & Buckley, 1991).

Naturally, when term dimensionality reduction is selected it must be carried out with care in order to remove only undesired terms which do not carry any semantic significance about the author's expressed original ideas.

Furthermore there is a variety of approaches for the inductive construction of a text classifier. In addition to hard classifiers (6.5) there are also soft classifiers which are typically elucidated as follows.

$$K : D \rightarrow [0,\ 1] \tag{6.6}$$

Soft classifiers are also another potential application of bidirectional fuzzy logic which is discussed in detail at a later section. Furthermore soft classifiers in order to

make a decision on the categorization of a document they employ a threshold which is elucidated as follows.

$$\text{if } K(d_i) \geq \tau \Rightarrow K : d_i \rightarrow T \tag{6.7}$$

$$\text{if } K(d_i) < \tau \Rightarrow K : d_i \rightarrow F \tag{6.8}$$

Another approach is embodied by the probabilistic classifiers which categorize documents according to the probability $P(c_j \mid \vec{d}_i)$ that a document represented by a vector $\vec{d}_i = \{w_{0i}, w_{1i}, w_{2i}, w_{3i} \ldots w_{mi}\}$ belongs to c_j. Probabilistic classifiers compute the aforementioned probability according to Bayes' theorem.

$$P(c_j \mid \vec{d}_i) = \frac{P(c_j) P(\vec{d}_i \mid c_j)}{P(\vec{d}_i)} \tag{6.9}$$

Where $P(\vec{d}_i)$ is the probability that a random document is represented by $\vec{d}_i$ and $P(c_j)$ is the probability that a random document belongs to c_j. Furthermore, if any randomly selected dyad of term dimensions from the document vector $\vec{d}_i$ are statistically independent then $P(\vec{d}_i \mid c_j)$ is elucidated as a probabilistic classifier also known as a Naïve Bayes classifier (Li & Jain, 1998) as follows.

$$P(\vec{d}_i \mid c_j) = \prod_{t=1}^{m} P(w_{ti} \mid c_j) \tag{6.10}$$

It is noteworthy to observe here that a supervised machine learning strategy called relevance feedback is also often employed by probabilistic search systems. A novel synthesis of bidirectional fuzzy logic and relevance feedback for machine learning is discussed in detail at a later section.

Another type of text classifier is based on a decision tree with internal nodes representing terms, outward branches representing term significance and leaves

representing categories. A decision tree classifier recursively tests for all the terms of vector d_i until a leaf node and therefore a categorization decision is reached.

Another type of text classifier is based on a disjunctive normal form rule with premise keywords denoting the presence or absence of terms in d_i and the clause denoting the categorization decision.

Another type of text classifier is based on the vector space model whereby a category is represented by a vector c_j which is a linear classifier according to the cosine similarity between the dyad of category and unseen document vectors.

Furthermore, in the vector space model Rocchio's method computes the linear classifier vector c_j as follows.

$$\vec{d}_i = \{w_{0i}, w_{1i}, w_{2i}, w_{3i} \ldots w_{mi}\} \tag{6.11}$$

$$\vec{c}_j = \{w_{0j}, w_{1j}, w_{2j}, w_{3j} \ldots w_{mj}\} \tag{6.12}$$

$$T_r = \{d_0, d_1, d_2, d_3 \ldots d_T\} \tag{6.13}$$

$$w_{xj} = \beta \cdot \sum_{d_i \in K_j^+} \frac{w_{xi}}{|K_j^+|} - \gamma \cdot \sum_{d_i \in K_j^-} \frac{w_{xi}}{|K_j^-|} \tag{6.14}$$

$$K_j^+ = \left\{ d_i \in T_r \mid \mathring{\Phi}(d_i, c_j) = T \right\} \tag{6.15}$$

$$K_j^- = \left\{ d_i \in T_r \mid \mathring{\Phi}(d_i, c_j) = F \right\} \tag{6.16}$$

Where T_r is the training set of paradigm documents, β, γ are control parameters of the impact of positive and negative term paradigms respectively and w_{xi} is the weight of term t_x in document d_i (Joachims, 1998).

Furthermore, another approach is to use neural networks as text classifiers whereby the input elements represent the terms, the output elements represent the categories and the weights on the edges interconnecting the elements of the neural network represent their associations.

The term weights of a document to be classified are entered into the input elements which are activated with a forward propagation through the network and the value of the output elements determines the category of the document.

It is not unusual to train neural networks with back propagation whereby the term weights of a document to be classified are entered into the input elements and if the output value indicates an inappropriate classification the error is back propagated in order to modify the network parameters so that the error is eliminated or at least reduced.

Furthermore non-linear neural networks contain a large number of layers of elements representing term associations of higher complexity. Another strategy for text classifiers is based on the k nearest neighbours algorithm (Yang, 1994).

Such a classifier examines whether the k training documents most similar to the d_i document under classification belong also to the same category. Hence if the majority of the k documents belong to the same category so does d_i and vice versa.

Another approach is based on support vector machines whereby a variety of surfaces are identified which separate the positive from the negative training paradigms and the decision surface which separates the positives from the negatives by the greatest possible distance is the decisive classifier.

Finally the boosting method relies on k classifiers which are obtained with the same learning method (Schapire & Singer, 2000). The k classifiers are trained sequentially considering the errors of the preceding classifiers in an attempt to eliminate similar future incorrect decisions in text categorisation.

6.5 Similarity heuristics

Statistical data of a variety of parameters including term occurrences, encapsulated context, syntax, proximity to exoteric synonymous lexicon from a thesaurus, esoteric and exoteric link structure analysis, form the bases of various document scoring methodologies for modern information retrieval systems (Brin & Page, 1998).

All these parameters are critical for different methodologies in order to compute the hierarchy of the search results. These document scoring techniques have two aims.

The first aim is to ascertain the significance of each document rank within the specified query-resulting document-collection.

The second aim is to reflect the credibility and importance of each document independently, utilizing metrics such as number of citations, quality of reference sources, number of readers (Brin & Page, 1998), etc.

In addition, the tetrad of the most popular documents similarity coefficients are listed subsequently, the inner product, cosine, dice and jaccard similarity coefficients (Ganesan, Garcia-Molina & Widom, 2003; Salton & McGill, 1997; Wen, Nie & Zhang, 2002).

For all the similarity coefficient formulae the reader is referred to Chapter 4.

All these similarity coefficients provide a numerical quantification of the association of two documents or the arithmetical expression of the homology between a document and a query.

Therefore, an information retrieval system can display retrieved documents in decreasing order of significance corresponding to their similarity with the query issued by the searcher.

6.6 Architectonic model and experimental data analysis

The architectonic model of the experimental information retrieval system *ANACALYPSE* is designed according to distributed object oriented architecture. A large number of objects are esoteric among a pair of exoteric entities which coordinate the interactions and communication of their esoteric collections of objects.

The pair of distinct exoteric entities are the server and the client(s) which can be arranged to operate in a parallel configuration in order for a large number of simultaneous clients to interact synchronously in parallel with a monadic server.

Each client consists of a variety of objects such as interactive dialogue and user input components, user historical data preservation entities, graphical user interfaces, output displays, bidirectional fuzzy relevance feedback controls, such that all the interactions between the experts and the machines are automated without the requirement for intervention by knowledge engineers.

The monadic server is designed in such a way that it can be either esoteric or exoteric to the machine(s) where the client(s) are hosted. Furthermore, the monadic server consists of a large number of objects for various operations.

The following Figure 6.3 portrays an overview of the architectonic model of the *ANACALYPSE* academic information retrieval system. Furthermore, Figure 6.4 portrays an overview of the Metagram matrix vector data structure and how it is constructed. In addition, Figure 6.5 displays a sample of a Metagram matrix vector data structure and sample similarities and statistics dictionary data structures.

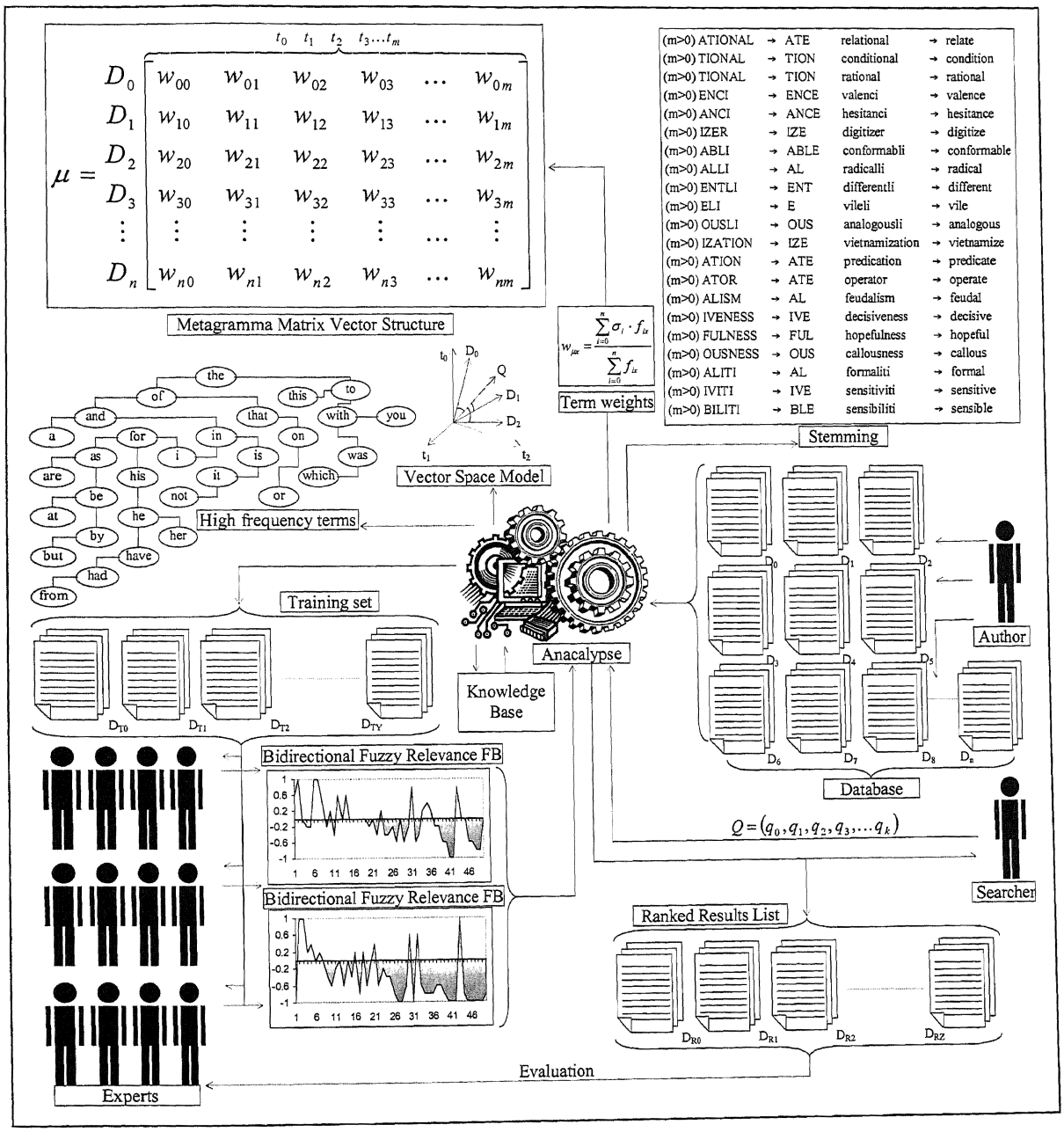

Figure 6.3. Graph of the architectonic model of the *ANACALYPSE* information retrieval system.

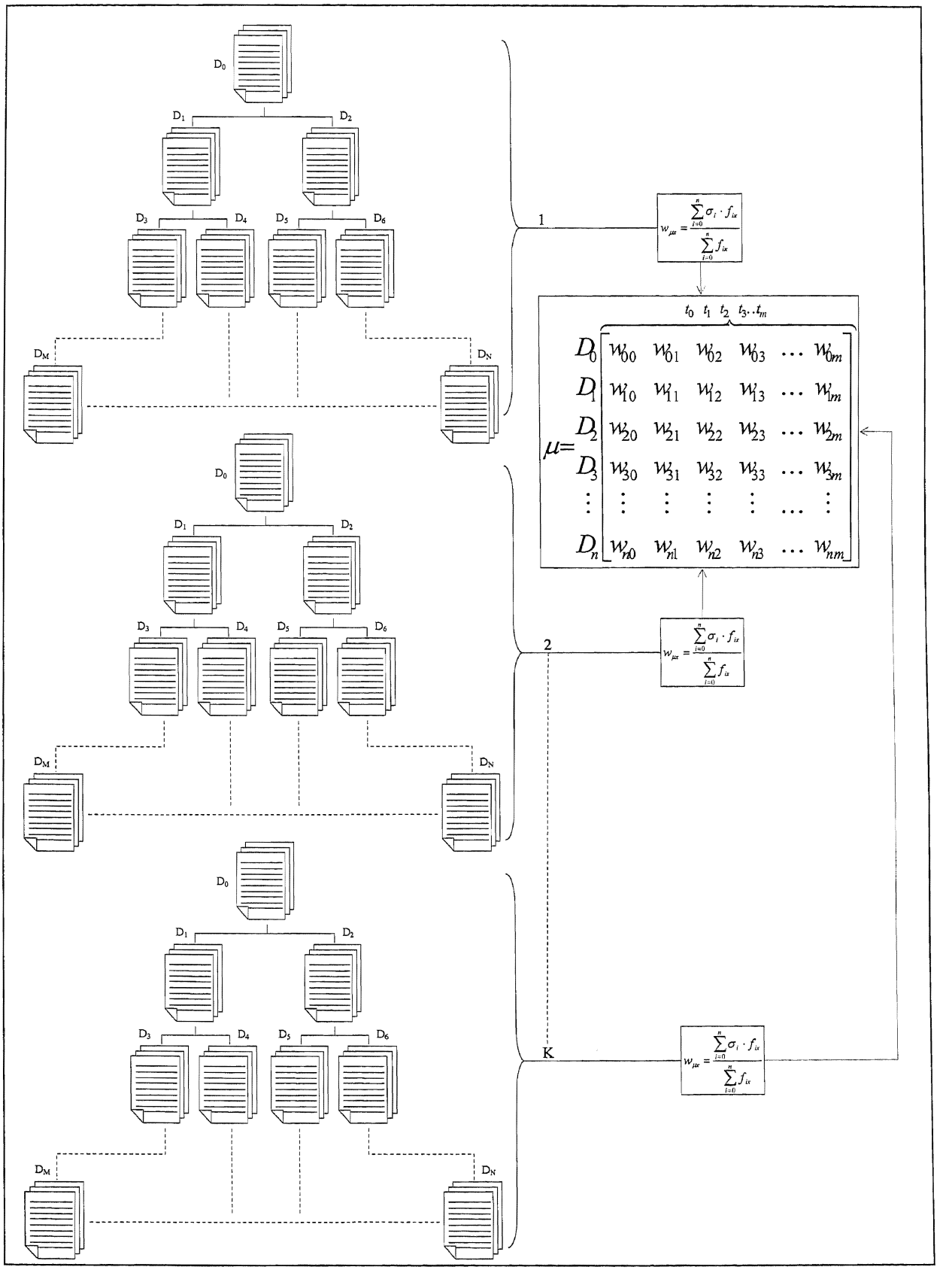

Figure 6.4. Graph of the *Metagram* matrix vector data structure.

```
metagram[doc] = [nobel, prize, medicine, research, publications, journals, alfred, libraries, organizations, institutes, nitroglycerin, ..... oxide]
metagram[06] = [0.98981099884851221, 0.98456786974728632, 0.97718483274489234, 0.86697340697583991, ..... 0.12859241637244601]
metagram[16] = [0.98546191149887772, 0.93646284072434476, 0.90135890926301165, 0.00000000000000000, ..... 0.19706830208761705]
metagram[21] = [0.94112007548919845, 0.92542580313232978, 0.87623912815927198, 0.82454039457154966, ..... 0.00000000000000000]
metagram[33] = [0.95423123211290122, 0.93111960547258299, 0.92430344721677821, 0.88693124153464321, ..... 0.08048007504074687]
metagram[37] = [0.90160573786804632, 0.88346564197036592, 0.82600964730899842, 0.82198053634239585, ..... 0.00000000000000000]
metagram[46] = [0.94665742408873154, 0.93712333793150773, 0.81933406674983811, 0.00000000000000000, ..... 0.07290988490226291]
metagram[57] = [0.92579547107221174, 0.90830770904202307, 0.87497236655428541, 0.81750854042513517, ..... 0.00000000000000000]
metagram[64] = [0.93651723143623578, 0.93196367643936018, 0.96743296017011353, 0.88362907206902683, ..... 0.16068248241937994]
metagram[79] = [0.95935847241844261, 0.92417958341527786, 0.91694406070371457, 0.00000000000000000, ..... 0.00000000000000000]
metagram[83] = [0.95598793735123628, 0.93752112062932955, 0.87156210581212026, 0.82836898090270008, ..... 0.22881593439579967]
metagram[101] = [0.95322666272989687, 0.87092852144335 42, 0.82614387318568749, 0.87732985878840808, ..... 0.00000000000000000]
metagram[102] = [0.97785545982754007, 0.90582799750245 02, 0.93031688172566296, 0.00000000000000000, ..... 0.00000000000000000]
metagram[103] = [0.95702186853519633, 0.95299905557213 67, 0.91599282418456163, 0.88873803717323474, ..... 0.00000000000000000]
metagram[104] = [0.95179208786656222, 0.91198043588675 78, 0.89750583703080338, 0.83850386608727128, ..... 0.00000000000000000]
        :          :                                  :                    :                    :                    :
        :          :                                  :                    :                    :                    :
metagram[120] = [0.89951845443846522, 0.86585164484989641, 0.8491383509084972, 0.82799540670631666, ..... 0.00000000000000000]


simDict[doc] = [ dotProduct, cosine, dice, jaccard, overlap ]
simDict[10] = [0.20679316874670264, 0.34890578798630029, 0.057132837018295572, 0.029406455627472863, 0.0101030566871706940]
simDict[70] = [0.19898266574024093, 0.33212442436153994, 0.041637535519733010, 0.021261403991820922, 0.0075024044601455743]
simDict[90] = [0.15687373954960274, 0.27362252273292875, 0.034871556320114648, 0.017745179167431124, 0.0057545641713443009]
simDict[75] = [0.21254674893606401, 0.24536569192934274, 0.022759848403893671, 0.011510917571404274, 0.0085577155317590192]
simDict[63] = [0.17274791787395433, 0.22383145511703845, 0.030484669140501425, 0.015478259378259260, 0.0091214311669042883]
simDict[47] = [0.16842100458155013, 0.21327402006106963, 0.027586608167009807, 0.013986220272705208, 0.0086379950686915966]
simDict[09] = [0.17636836268974102, 0.21074526582761854, 0.028127140460233870, 0.014264175463522901, 0.0098939192384868875]
simDict[28] = [0.15041674260170118, 0.20622582480090035, 0.030942098219083183, 0.015714163707983175, 0.0082773463274213501]
simDict[25] = [0.21269147074966502, 0.20518330300402179, 0.017539642762967683, 0.008847411600912309, 0.0094406615249038301]
simDict[39] = [0.18591539078978575, 0.20341355372640588, 0.013538444248254729, 0.006815356788081063, 0.0056609775389114469]

d1=cosRank-expertRank, d2=googleRank-expertRank, d3=dotProductRank-expertRank,
d4=diceRank-expertRank, d5=jaccardRank-expertRank, d6=overlapRank-expertRank
statsDict[doc] = [googleRank, cosRank, expertRank, d1*d1, d2*d2, dotProductRank,
                  diceRank, jaccardRank, d3*d3, d4*d4, d5*d5, overlapRank, d6*d6]
statsDict[63] = [7.0, 5.0, 6.0, 1.0, 1.0, 7.0, 5.0, 5.0, 1.0, 1.0, 1.0, 4.0, 4.0]
statsDict[70] = [8.0, 2.0, 2.0, 0.0, 36.0, 4.0, 2.0, 2.0, 4.0, 0.0, 0.0, 8.0, 36.0]
statsDict[25] = [3.0, 9.0, 9.0, 0.0, 36.0, 1.0, 9.0, 9.0, 64.0, 0.0, 0.0, 3.0, 36.0]
statsDict[09] = [1.0, 7.0, 7.0, 0.0, 36.0, 6.0, 6.0, 6.0, 1.0, 1.0, 1.0, 2.0, 25.0]
statsDict[10] = [2.0, 1.0, 1.0, 0.0, 1.0, 3.0, 1.0, 1.0, 4.0, 0.0, 0.0, 1.0, 0.0]
statsDict[75] = [9.0, 4.0, 5.0, 1.0, 16.0, 2.0, 8.0, 8.0, 9.0, 9.0, 9.0, 6.0, 1.0]
statsDict[28] = [4.0, 8.0, 4.0, 16.0, 0.0, 10.0, 4.0, 4.0, 36.0, 0.0, 0.0, 7.0, 9.0]
statsDict[90] = [10.0, 3.0, 3.0, 0.0, 49.0, 9.0, 3.0, 3.0, 36.0, 0.0, 0.0, 9.0, 36.0]
statsDict[39] = [5.0, 10.0, 10.0, 0.0, 25.0, 5.0, 10.0, 10.0, 25.0, 0.0, 0.0, 10.0, 0.0]
statsDict[47] = [6.0, 6.0, 8.0, 4.0, 4.0, 8.0, 7.0, 7.0, 0.0, 1.0, 1.0, 5.0, 9.0]

n=10.0
statsDict[SpearmanCos]          = 1.0-{6.0*sum(d1*d1)/[n*(n*n-1.0)]}
statsDict[SpearmanCos]          = 0.866666666667
statsDict[SpearmanGoogle]       = 1.0-{6.0*sum(d2*d2)/[n*(n*n-1.0)]}
statsDict[SpearmanGoogle]       = -0.236363636364
statsDict[SpearmanDotProduct]   = 1.0-{6.0*sum(d3*d3)/[n*(n*n-1.0)]}
statsDict[SpearmanDotProduct]   = -0.0909090909091
statsDict[SpearmanDice]         = 1.0-{6.0*sum(d4*d4)/[n*(n*n-1.0)]}
statsDict[SpearmanDice]         = 0.927272727273
statsDict[SpearmanJaccard]      = 1.0-{6.0*sum(d5*d5)/[n*(n*n-1.0)]}
statsDict[SpearmanJaccard]      = 0.927272727273
statsDict[SpearmanOverlap]      = 1.0-{6.0*sum(d6*d6)/[n*(n*n-1.0)]}
statsDict[SpearmanOverlap]      = 0.0545454545455
```

Figure 6.5. Excerpt of a *Metagram* matrix vector data structure, sample similarities and statistics dictionary data structures.

For instance, a few of these operations include the following activities. Locating and transferring from distributed machines a large number of relevant documents according to the searcher's request.

Performing the metamorphosis of tag bearing text to pure text, whilst preserving the text descriptive-fields of hyper media objects such as images, acoustic and cinematographic objects.

Identifying and extracting undesired high frequency terms from the document. Discovering and preserving the types and roots of words. Esoterically preserving the machine's historical data, by tracking all the documents visited by the searcher starting from a root page and traversing down a logical tree.

Performing the analysis and automatic metamorphosis of all documents into vectors. Automatically presenting the machine training sample of documents to the experts for reading and automatically eliciting their bidirectional fuzzy relevance feedback based on the event activated human computer interaction model.

Performing the automatic syntheses of experts' bidirectional fuzzy relevance feedback with the machine training sample of corresponding document vectors in a new Metagram matrix vector structure.

Performing the automatic computation of the term weights of the Metagram matrix vector structure, according to the experts' bidirectional fuzzy relevance feedback on the machine training sample of documents. Performing the automatic computation of

the similarities of the Metagram matrix vector structure with all the term vectors of all the documents which have not been read or seen by the experts, etc.

Also, for evaluation purposes another object in the server selects a uniform random sample of documents from the ranked results list for inspection by the experts. The Wichmann Hill multiplicative congruential generator with a period of 6,953,607,871,644 is employed to create a series of random numbers which are applied to the selection of the evaluation sample of documents.

The Wichmann Hill multiplicative congruential generator (L'Ecuyer, 1988; 1990) is one of the most popular random number generators and it utilizes the following recurrence formula.

$$x_i = \alpha \cdot x_{i-1} \bmod m \tag{6.17}$$

The parameters α and m m are elucidated such that $1 < \alpha < m$, the initial seed is x_0 such that $0 < x_0 < m$ and the desired output sequence of random numbers is elucidated as follows.

$$u_i = \frac{x_i}{m} \tag{6.18}$$

Typically m is approximately the largest integer that the computer's arithmetic processing unit can exactly represent using single precision. It is not unusual to select parameter values according to theoretical and empirical considerations for both α and m, for instance if m is a prime number then the system attains maximum possible cycle length, 1…m-1 are generated before any are repeated.

Thus for 32 bit word machines it is not unusual for m to take the value 2^{31}-1 and for α to take the value 16807 (Bratley, Fox & Schrage, 1983; Wichmann & Hill, 1982), naturally (6.17) yields results appreciably faster when computed in modern machines with double precision arithmetic units and wider mathematical registers instead of single precision.

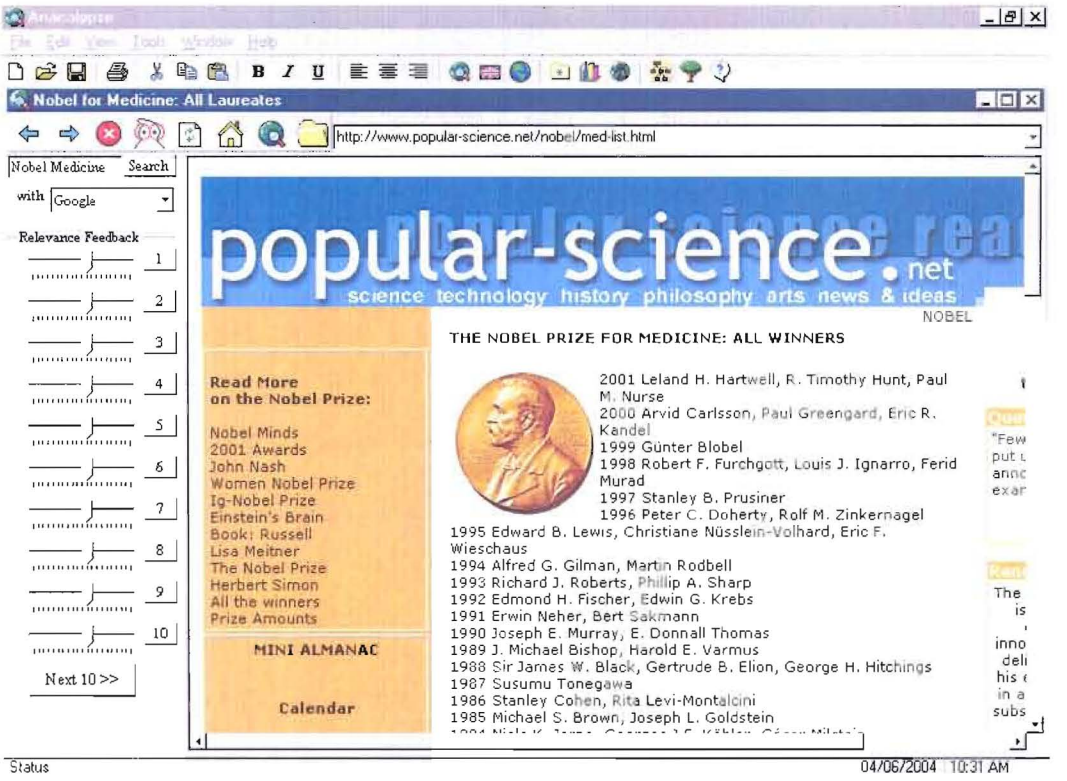

Figure 6.6. The first stage of a document downloaded from the internet.

This Figure 6.6 portrays the first stage of a document downloaded from the internet. After the selection of the evaluation sample of documents from the ranked results list is complete another object in the server transfers the evaluation sample of documents to an object in the client for expert inspection and ranking evaluation.

This object in the client automatically presents the evaluation sample of documents to the experts for reading and based on the event activated human computer interaction model automatically elicits the expert's bidirectional fuzzy relevance feedback and returns it to the originating object in the server.

The designated object in the server computes the relative ranks of both *ANACALYPSE* and Google, compared to the expert ranking evaluation and computes all the corresponding Spearman correlation coefficients.

In synopsis, on document relevance issues of information retrieval a series of experiments can be conducted with interactive expert supervised relevance feedback in order to rank the retrieved information from irrelevant to highly relevant and to rank the retrieved information having negative impact or positive impact thus the group impact for a particular system can then be precisely computed.

```
<p align="center">Read also:<a href="http://www.amazon.com/exec/obidos/ASIN/0806520256/popularscience0c"><b><br>
</b><img src="http://images.amazon.com/images/P/0806520256.01.jpg" width=59 height=90 vspace=3 align=top hspace=5
border=0><b>
<br>
Nobel Prize Women <br>
in Science :<br>
</b></a><a href="http://www.amazon.com/exec/obidos/ASIN/0806520256/popularscience0c"><b>Their
Lives, Struggles, and Momentous Discoveries</b></a> <br>
by Sharon Bertsch McGrayne <br>
<br>
</p>
<p> </p>
</td>
<td width="10"> </td>
<td width="405">
<table width="405" border="0" cellspacing="0" cellpadding="0">
<tr>
<td align="left" valign="top">
<table width=100% border=0 cellpadding=4 cellspacing=0 bgcolor=#8ca5ce>
<tr>
<td bgcolor=#ffffff valign=top width=100%>
<p><b>THE NOBEL PRIZE FOR MEDICINE: ALL WINNERS<br>
</b> </p>
<p> <a href="/nobel/theprize.html"><img src="/img/nobel_medal.jpg" alt="This is an image of the gold Nobel medal
awarded with a personal diploma and a prize amount, in memory of Alfred Nobel, founder of the prestigious Nobel
Prize. The Nobel prize is the famous international award given yearly since 1901 for achievements in physics,
chemistry, medicine, literature and peace. After 1968 the Nobel prize is awarded also for achievements in economics.
The lack of a Nobel prize for mathematics was the object of many discussions. The truth is probably that Alfred Nobel
was more interested in applied science, and viewed Mathematics as a more abstract field. However mathematicians
and computer scientists have now their own prestigious awards, respectively the famous Fields medal and the Turing
award."
width="109" height="112" align="left" border="0"></a>
2001 Leland H. Hartwell, R. Timothy Hunt, Paul M. Nurse<br>
2000 Arvid Carlsson, Paul Greengard, Eric R. Kandel <br>
1999 Günter Blobel <br>
1998 Robert F. Furchgott, Louis J. Ignarro, Ferid Murad <br>
1997 Stanley B. Prusiner <br>
1996 Peter C. Doherty, Rolf M. Zinkernagel <br>
1995 Edward B. Lewis, Christiane Nüsslein-Volhard, Eric F. Wieschaus <br>
1994 Alfred G. Gilman, Martin Rodbell <br>
1993 Richard J. Roberts, Phillip A. Sharp <br>
1992 Edmond H. Fischer, Edwin G. Krebs <br>
1991 Erwin Neher, Bert Sakmann <br>
1990 Joseph E. Murray, E. Donnall Thomas <br>
1989 J. Michael Bishop, Harold E. Varmus <br>
1988 Sir James W. Black, Gertrude B. Elion, George H. Hitchings<br>
1987 Susumu Tonegawa <br>
1986 Stanley Cohen, Rita Levi-Montalcini <br>
1985 Michael S. Brown, Joseph L. Goldstein <br>
1984 Niels K. Jerne, Georges J.F. Köhler, César Milstein <br>
1983 Barbara McClintock <br>
1982 Sune K. Bergström, Bengt I. Samuelsson, John R. Vane <br>
```

Figure 6.7. Excerpt of the raw tagged data of a downloaded document.

This Figure 6.7 portrays an excerpt of the raw tagged data of a downloaded pre-processed document. The expert supervised training of the bidirectional fuzzy information retrieval system can be conducted in order to eventually achieve unsupervised system document relevance estimation based on the initial small training

sample and a relevance function which is formed through syntheses of the supervised training, the document similarity measure and the term weighting function.

Read also:
(image)
Nobel Prize Women
in Science :
[61]
Their Lives, Struggles, and Momentous Discoveries
[62]
by Sharon Bertsch McGrayne

THE NOBEL PRIZE FOR MEDICINE: ALL WINNERS

This is an image of the gold Nobel medal awarded with a personal diploma and a prize amount, in memory of Alfred Nobel, founder of the prestigious Nobel Prize. The Nobel prize is the famous international award given yearly since 1901 for achievements in physics, chemistry, medicine, literature and peace. After 1968 the Nobel prize is awarded also for achievements in economics. The lack of a Nobel prize for mathematics was the object of many discussions. The truth is probably that Alfred Nobel was more interested in applied science, and viewed Mathematics as a more abstract field. However mathematicians and computer scientists have now their own prestigious awards, respectively the famous Fields medal and the Turing award.
[63]
2001 Leland H. Hartwell, R. Timothy Hunt, Paul M. Nurse
2000 Arvid Carlsson, Paul Greengard, Eric R. Kandel
1999 Günter Blobel
1998 Robert F. Furchgott, Louis J. Ignarro, Ferid Murad
1997 Stanley B. Prusiner
1996 Peter C. Doherty, Rolf M. Zinkernagel
1995 Edward B. Lewis, Christiane Nüsslein-Volhard, Eric F. Wieschaus
1994 Alfred G. Gilman, Martin Rodbell
1993 Richard J. Roberts, Phillip A. Sharp
1992 Edmond H. Fischer, Edwin G. Krebs
1991 Erwin Neher, Bert Sakmann
1990 Joseph E. Murray, E. Donnall Thomas
1989 J. Michael Bishop, Harold E. Varmus
1988 Sir James W. Black, Gertrude B. Elion, George H. Hitchings
1987 Susumu Tonegawa
1986 Stanley Cohen, Rita Levi-Montalcini
1985 Michael S. Brown, Joseph L. Goldstein
1984 Niels K. Jerne, Georges J.F. Köhler, César Milstein
1983 Barbara McClintock
1982 Sune K. Bergström, Bengt I. Samuelsson, John R. Vane

Figure 6.8. Metamorphosis of the raw tagged data into pure text by the *ANACALYPSE* system.

This Figure 6.8 portrays an excerpt of the text extracted from a hypertext document. This is the first processing stage of the raw tagged data of a downloaded document. The bidirectional fuzzy interval is [-1, 1] and the documents which score -1 have high negative impact in the antithesis pole, the documents that score 0 are neutral and the documents that score 1 are highly relevant in the thesis pole.

The gradient from one stage to another can be further attributed a linguistic characterization which would illustrate the inclination from one stage to the next such as, possibly relevant, applicable, pertinent, important, etc.

read also:
(image)
nobel prize women
science :
[61]
lives, struggles, momentous discoveries
[62]
sharon bertsch mcgrayne

nobel prize medicine: winners

image gold nobel medal awarded personal diploma prize amount, memory alfred nobel, founder prestigious nobel prize. nobel prize famous international award given yearly 1901 achievements physics, chemistry, medicine, literature peace. 1968 nobel prize awarded achievements economics. lack nobel prize mathematics object discussions. truth probably alfred nobel interested applied science, viewed mathematics abstract field. mathematicians computer scientists prestigious awards, respectively famous fields medal turing award.
[63]
2001 leland h. hartwell, r. timothy hunt, paul m. nurse
2000 arvid carlsson, paul greengard, eric r. kandel
1999 günter blobel
1998 robert f. furchgott, louis j. ignarro, ferid murad
1997 stanley b. prusiner 1996 peter c. doherty, rolf m. zinkernagel
1995 edward b. lewis, christiane nüsslein-volhard, eric f. wieschaus
1994 alfred g. gilman, martin rodbell
1993 richard j. roberts, phillip a. sharp
1992 edmond h. fischer, edwin g. krebs
1991 erwin neher, bert sakmann
1990 joseph e. murray, e. donnall thomas
1989 j. michael bishop, harold e. varmus
1988 sir james w. black, gertrude b. elion, george h. hitchings
1987 susumu tonegawa
1986 stanley cohen, rita levi-montalcini
1985 michael s. brown, joseph l. goldstein
1984 niels k. jerne, georges j.f. köhler, césar milstein
1983 barbara mcclintock
1982 sune k. bergström, bengt i. samuelsson, john r. vane

Figure 6.9. *ANACALYPSE* removes stop words from the pure text document.

This Figure 6.9 portrays an excerpt of the text after stop-words removal. This is the second processing stage of the raw tagged data of a downloaded document. The queries are selected by the experts in their areas of expertise in order for the relevance training to be accurate.

Furthermore, a series of experiments have been conducted with the elicited interactive expert supervised relevance feedback with a dyad of principal objectives, to automatically compute the relevance of the unseen documents according to the bidirectional fuzzy relevance feedback given by experts on a small sample of seen documents and to measure the relative effectiveness of the experimental and commercial information retrieval systems compared to the experts' rankings.

read also:
(image)
nobel prize women
scienc :
[61]
lives, struggles, moment discoveri
[62]
sharon bertsch mcgrayn

nobel prize medicine: winner

imag gold nobel medal award person diploma prize amount, memori alfred nobel, founder prestigi nobel prize. nobel prize famou intern award given yearli 1901 achiev physics, chemistry, medicine, literatur peace. 1968 nobel prize award achiev economics. lack nobel prize mathemat object discussions. truth probabl alfr nobel interest appli science, view mathemat abstract field. mathematician comput scientist prestigi awards, respect famou field medal ture award.
[63]
2001 leland h. hartwell, r. timothy hunt, paul m. nurse
2000 arvid carlsson, paul greengard, eric r. kandel
1999 günter blobel
1998 robert f. furchgott, louis j. ignarro, ferid murad
1997 stanley b. prusiner 1996 peter c. doherty, rolf m. zinkernagel
1995 edward b. lewis, christiane nüsslein-volhard, eric f. wieschaus
1994 alfred g. gilman, martin rodbell
1993 richard j. roberts, phillip a. sharp
1992 edmond h. fischer, edwin g. krebs
1991 erwin neher, bert sakmann
1990 joseph e. murray, e. donnall thomas
1989 j. michael bishop, harold e. varmus
1988 sir james w. black, gertrude b. elion, george h. hitchings
1987 susumu tonegawa
1986 stanley cohen, rita levi-montalcini
1985 michael s. brown, joseph l. goldstein
1984 niels k. jerne, georges j.f. köhler, césar milstein
1983 barbara mcclintock
1982 sune k. bergström, bengt i. samuelsson, john r. vane

Figure 6.10. After stop word removal *ANACALYPSE* performs stemming on the pure text document.

This Figure 6.10 portrays an excerpt of the text after the process of stemming is complete. This is the third processing stage of the raw tagged data of a downloaded

document. The propaedeutic training sets consist of syntheses of the sample documents and their corresponding relevance feedback values.

Subsequently the experimental system automatically computes the representation of the training sets by term vectors weighted according to syntheses of expert bidirectional fuzzy relevance feedback associated with the subject matter of the sample documents and normalized frequency weighting function values.

Subsequently the training sets are automatically synthesized into a monadic matrix vector structure called Metagram, which is preserved and used for vector comparison with the unseen document vectors. The comparison of the monadic matrix vector structure Metagram with the unseen document vectors is accomplished through vector analysis and similarity computation of the cosine between each dyad of vectors. All this information is stored and used to generate a knowledge base for further information retrieval sessions.

Table 6.1. Heterogeneous user population distribution selected for the information retrieval experiments.

User ID	Gender	Age	Native English
1	1	0	1
2	1	0	0
3	1	1	0
4	1	1	1
5	0	1	1
6	0	0	1
7	0	0	0
8	0	1	0

The experts listed in Table 6.1 are researchers from the University of Luton and are selected in a diverse approach so that they represent the whole heterogeneous

spectrum of gender, age and origin characteristics. The gender parameter corresponds to male if one is indicated and female if zero is indicated.

The age parameter corresponds to greater or equal to thirty if one is indicated and younger than thirty if zero is indicated. The native English parameter corresponds to English native speaker if one is indicated and non English native if zero is indicated.

Furthermore, the term weighting approaches in *ANACALYPSE* incorporate a vector length normalization factor. Numerous strategies have been reported regarding the elicitation of relevance information and the representation of meaning (Freeman, 2000; Kobayashi, Chang, & Sugeno, 2002).

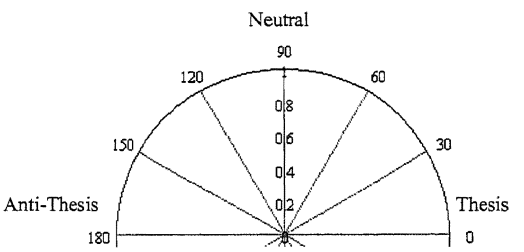

Figure 6.11. Hemi-cyclical bidirectional fuzzy relevance feedback and document similarity diagram.

Herein a novel approach is developed whereby expert bidirectional fuzzy relevance feedback is elucidated on the basis of the hemi-cyclical thesis-antithesis methodology portrayed in Figure 6.11. The thesis represents a specific view and the antithesis represents the opposite view of the thesis. All the rest of the results are somewhere in between these two extreme locations in various angular displacements, with the majority being neutral vectors occupying the area around the middle position.

At this point, it is noteworthy to remark that in a dyadic antithesis like the aforementioned example, the north and south poles are not involved. Also, note that the cyclical diagram serves not only as a bidirectional fuzzy relevance feedback map but also as a document similarity chart as shown in Figure 6.11.

That is, while the antithesis location carries a -1 relevance feedback weight, the thesis location carries a 1 and as the angle of the unseen document vectors decreases towards either extreme location, the similarity between the unseen documents and the seen documents represented by either extreme location increases.

Furthermore, according to Salton's cosine measure, one can easily see that although there is clearly an antithesis between the –x and x positions of the cyclical diagram, the y and –y positions are left unaffected since they have the same cosine value.

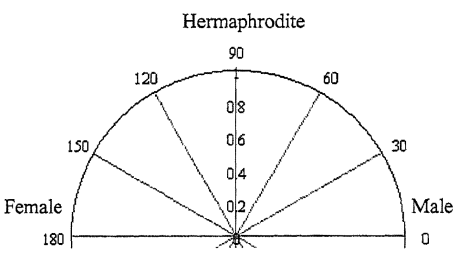

Figure 6.12. A thesis-antithesis example of hemi-cyclic bidirectional fuzzy relevance and document similarity.

To illustrate this concept, let us consider an example as shown in Figure 6.12. Consider a medical database with a variety of various biological traits of creatures from the animal kingdom. For instance, if the expert issues a query such as "biological traits of male animals", then the most relevant information retrieval result

in the thesis location (i.e. the document vector location) might be male; at the antithesis location might be female, at the neutral location might be hermaphrodite.

At this point, it is noteworthy to remark that in a dyadic antithesis, the north and south poles are not involved. Also, note that the hemi-cyclical diagram serves not only as a bidirectional fuzzy relevance feedback map but also as a document similarity chart.

That is, while the antithesis location carries a -1 relevance feedback weight, the thesis location carries a 1 and, as the angle of the unseen document vectors decreases towards either extreme location, the similarity between the unseen document and the seen document represented by either extreme locations increases.

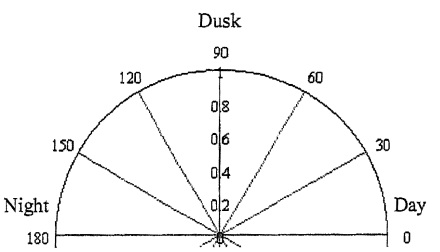

Figure 6.13. A dyadic antitheses example of hemi-cyclic bidirectional fuzzy relevance and document similarity.

In order to illustrate this concept, let us consider Figure 6.13 which portrays another concise example. Furthermore, according to Salton's cosine measure, one can easily see that although there is clearly an antithesis between the west –x and east x positions of the hemi-cyclical diagram, the north y and south –y positions are left unaffected since they have the same cosine value.

6.7 Statistical analysis of experimental data

The experimental testing took place in the University of Luton with 8 volunteer subjects. The profile of these users is shown in Table 6.1, with respect to age (4 under 30 years of age, 4 of 30 or over), gender (4 male, 4 female), native-English status (4 native English speakers, 4 whose native language was not English).

All the following combinations of similarity functions [ABFGH]A[NABCDE]-ABB-BBB were considered. The data are tabulated in Appendix B. Each user performed four searches, to cover the four possible combinations of stemming (on or off) and stop-word removal (on or off).

6.7.1 Descriptive Statistics

Owing to the balanced design, the mean value for all binary factors (Age, Gender, Native-English status, Stemming and Stop-word Removal) was 0.5 (50%). Basic statistics for Files History (number of pages visited during roaming) are given in Table 6.2 which follows.

Table 6.2. Basic Statistics for Files History.

Files History		
N	Valid	768
	Missing	0
Mean		2.73
Median		2.00
Std. Deviation		2.85
Percentiles	25	1.00
	50	2.00
	75	4.00

This Table 6.2 shows that the mean number of pages visited during roaming was 2.73 and 2 was the median. Basic statistics for the response variables are given in Table 6.3 which follows.

Table 6.3. Basic Statistics for Response Variables.

		GOOGLE	COSINE	DotProduct	JACCARD	OVERLAP
N	Valid	768	768	768	768	768
	Missing	0	0	0	0	0
Mean		6.6351E-02	.603835	.106691919	.611111	9.92898E-02
Median		8.4849E-02	.709091	6.66666667E-02	.793939	.103030
Std. Deviation		.3294	.360539	.373239177	.377911	.378027
Percentiles	25	-0.1849	.309091	-.151515152	.372727	-.151515
	50	8.4849E-02	.709091	6.66666667E-02	.793939	.103030
	75	.3545	.951515	.406060606	.890909	.345455

This Table 6.3 shows, for example, that the mean rank correlation between the Google rankings and the user's subsequently expressed relevance feedback scores was close to zero (0.066351). The mean rank correlation between *ANACALYPSE* rankings formed using the Cosine similarity formula and the user's relevance feedback, on the other hand, was 0.603835.

All five response variables were subjected to a Kolmogorov-Smirnov one-sample test in comparison with the Normal distribution. They all failed it: Overlap at the 5% significance level ($p = 0.026$), and the other four variables at the 1% significance level. Thus it was concluded that none of these variables should be treated as coming from a Normal (Gaussian) distribution; hence that non-parametric tests should be used in comparing them.

6.7.2 Comparisons among Response Variables

Taking the median as a more robust measure of central tendency in the present context, these five response variables can be ranked in Table 6.4 as follows in terms of their correlations with the user's expressed ratings.

Table 6.4. Ordering of Response Variables.

Position	Variable	Median Rank Correlation with User Relevance Ratings
1.	Jaccard	0.7939
2.	Cosine	0.7091
3.	Overlap	0.1030
4.	Google	0.0848
5.	DotProduct	0.0667

Google does not make use of information from the use's feedback on the first batch of pages viewed to rate the relevance of the second batch; the other four variables do. Three of the four methods making use of this additional information "beat" Google, according to this ordering.

Two questions arise naturally in this respect: (1) do the four similarity measures differ significantly among themselves? and (2) do they give better results, in terms of matching the user's expressed preferences, than Google?

To shed light on the first question, a Friedman test (a non-parametric equivalent to the 1-way analysis of variance) was performed. The mean ranks of these four variables (ranked between themselves) gave the following ordering: Cosine (3.27), Jaccard (3.13), DotProduct (1.84), Overlap (1.76).

Paired comparisons, using the Wilcoxon signed rank test showed that the difference between the middle two (Jaccard and DotProduct) was highly significant ($Z = 20.21$, $p < 0.0005$), whereas the differences between the first two (Cosine and Jaccard) and the last two (DotProduct and Overlap) were not. In summary: Cosine and Jaccard methods perform roughly equally well and significantly better than either DotProduct or Overlap.

With regard to the second question -- whether the methods that employ user feedback do better than Google (which does not) -- a Wilcoxon signed ranks test was performed to compare Overlap (the worst of the four feedback-using methods) with Google.

This gave a Z-score of 2.43 ($p = 0.015$). Thus, even the worst of the four feedback-using methods gives a stronger association with the user's final ratings than Google, the commercial system.

Thus it has been demonstrated that **a payoff does exist** for the work involved in expressing feedback for the first batch of documents: if the user is prepared to put in work by rating a small number of seen documents, they can realistically expect better ranking among the large number of documents that they have not seen than would otherwise be the case.

To emphasize this point, the variables CminusG (difference between Cosine and Google scores) and JminusG (difference between Jaccard and Google scores) were calculated. JminusG failed the Kolmogorov test for normality but CminusG,

fortuitously, passed it. A histogram of its distribution is shown in Figure 6.14 which follows.

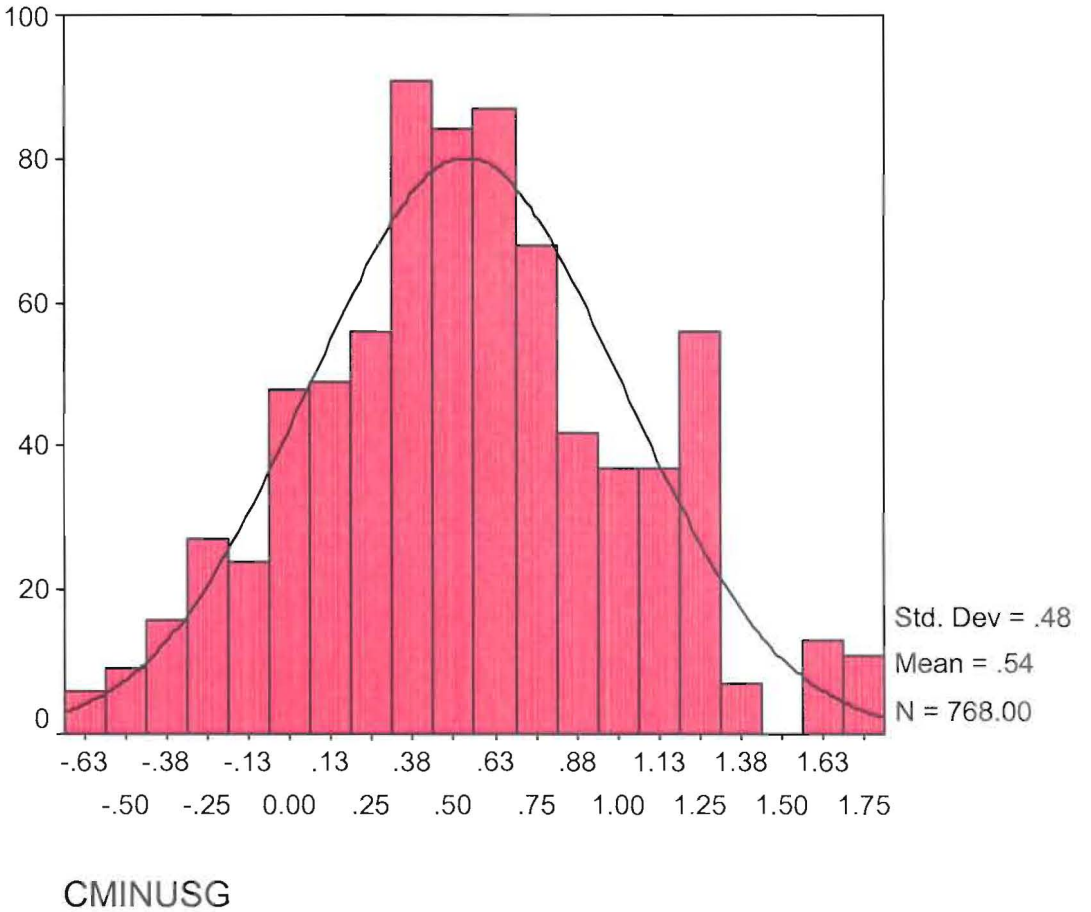

Figure 6.14. A histogram of the distribution of the *CMINUSG* variable.

It can be seen that the mean difference between the Cosine-based correlation and the Google-based correlation is 0.54. As this variable, *CMINUSG*, is approximately normally distributed a t-test was performed, showing (please see Table 6.5) that its mean is very significantly different from zero.

Indeed, 85.9% of the values are positive, meaning that in only 14.1% of the comparisons does the Google correlation exceed the Cosine-based *ANACALYPSE* correlation.

Table 6.5. Comparison of *Cosine* with Google Scores.

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
CMINUSG	768	.5375	.4771	1.722E-02

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
CMINUSG	31.220	767	.000	.5375	.5037	.5713

At this point it is noteworthy to re-iterate: Metagrams matter!

6.7.3 Effect of User Factors

Assuming henceforward that the Cosine and Jaccard methods are viable in this sort of application (and that DotProduct and Overlap methods can be "deprecated"), the effect of the user factors (Age, gender, English-status) were tested on these two viable response measures. The results, using the Mann-Whitney non-parametric test, are summarized in Table 6.6 which follows.

Table 6.6. Effects of User Factors on Response Variables.

Test Statistics

	COSINE	JACCARD
Mann-Whitney U	64513.500	72173.500
Wilcoxon W	138433.500	146093.500
Z	-3.004	-.506
Asymp. Sig. (2-tailed)	.003	.613

a Grouping Variable: AGE

Test Statistics

	COSINE	JACCARD
Mann-Whitney U	57823.500	42845.500
Wilcoxon W	131743.500	116765.500
Z	-5.184	-10.051
Asymp. Sig. (2-tailed)	.000	.000

a Grouping Variable: GENDER

Test Statistics

	COSINE	JACCARD
Mann-Whitney U	71809.000	68297.500
Wilcoxon W	145729.000	142217.500
Z	-.626	-1.767
Asymp. Sig. (2-tailed)	.532	.077

a Grouping Variable: Native English

These results show that Age has a significant effect when using the Cosine method (younger subjects getting higher scores) but not when using Jaccard. As regards Gender, this has a significant effect using both response variables (males getting higher scores). Native-English speaker status has no significant effect on either response variable.

This is not an ideal result. It implies that there is no outright "winning" similarity scoring method for all users but rather that the preferred similarity scoring method may depend on the age of the user. It also suggests that the current system is more satisfactory for males than females. Because of the small number of users involved (8

people for every experiment) these findings should be treated only as hints for future follow-up studies.

6.7.4 Effect of Experimental Manipulations

To throw some light on the effect of the experimental manipulations, two multiple linear regressions were performed using SPSS -- one with Cosine as the dependent variable, the other with Jaccard as the dependent variable.

Strictly speaking, this is not justified, as neither of the dependent variables can reasonably be assumed to come from a Normal distribution. However, in this case the objective is not to derive a valid predictive formula.

The objective in this instance is to use the SPSS Stepwise Regression procedure as a heuristic to indicate which experimental variables, in combination, have the most nearly linear relationship with the chosen response variables. The results are summarized in Table 6.7 which follows.

Table 6.7. Regression Coefficients predicting *Cosine* and *Jaccard*.

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.749	.017		44.453	.000
	REMOVAL	-.290	.024	-.403	-12.186	.000
2	(Constant)	.712	.020		35.186	.000
	REMOVAL	-.290	.024	-.403	-12.262	.000
	Files History	1.355E-02	.004	.107	3.261	.001

a Dependent Variable: COSINE

Coefficients

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.572	.019		30.408	.000
	Files History	1.445E-02	.005	.109	3.036	.002

a Dependent Variable: JACCARD

These results show that the Files-History variable (number of pages consulted during browsing) has a positive association with both response variables.

These results also show that the binary variable Removal (whether stop-word removal was in effect) has a strong negative effect on the Cosine score -- though not on Jaccard.

The Stemming variable was not included by the stepwise process in either case: it had negligible effect.

Thus, roughly speaking, we can assert that stemming is a waste of time and that stop-word removal (which is practiced very generally in the field of IR) is sometimes counter-productive.

On the other hand, using the files which a user visits from a particular root page, as well as the root page itself, to form the Metagram does seem to be worth the effort.

6.7.5 Summary of Findings

The findings of this analysis can be summed up briefly as follows.

- The *ANACALYPSE* technique of using feedback from seen documents to estimate the relevance of unseen documents does indeed result in better correspondence between system ranking of documents and users' preferences than the (feedback-free) ranking of a leading-edge commercial system, namely Google. This is the **key finding** of the present study.

- For the purpose of scoring the similarity between an *ANACALYPSE* Metagram and an unseen document, the Cosine and Jaccard similarity scoring methods give significantly better results than the DotProduct and Overlap methods.

- There appears to be an interaction between the similarity scoring method used and certain user characteristics. In particular, younger subjects seem to be more satisfied with the results of the Cosine method than the Jaccard method.

- It appears that males are more satisfied with the results of the *ANACALYPSE* process than females.

- Tracking the documents visited from a root page does give better results than simply using the root page to form a Metagram.

- Stemming does not seem to be worth the trouble.

- Stop-word removal, which is almost a universal norm in information retrieval, does not seem to be beneficial and can sometimes be harmful.

CHAPTER 7 – Conclusions

7.1 Introduction

This chapter presents conclusions drawn from the experimental data produced by the academic information retrieval system after experimentation procedures, while comparing the academic system to a commercial system and at the same time comparing results to the expert judgments.

7.2 Brief outcomes

In this section some brief outcomes are presented from the work of this thesis with regards to the information retrieval effectiveness using the proposed novel methodologies and the new academic information retrieval system.

- The primary aim of this research is to investigate methods of improving the effectiveness of current information retrieval systems. The introduced methods of using bidirectional relevance feedback from seen documents to estimate the relevance of unseen documents does indeed improve the effectiveness of the information retrieval process. This is the **key finding** of the present study.

- A foundational objective is to introduce a novel bidirectional, symmetrical fuzzy logic theory. This new theory is introduced herein as a valuable enhancement to the information retrieval process, including internet searches of distributed data objects.

- A further objective is to design, implement and apply the novel theory to an experimental information retrieval system. Herein, such a system, called *ANACALYPSE,* is implemented and automatically computes the relevance of a large number of unseen documents from expert relevance feedback on a small number of documents read.

- Another objective is to define a methodology used in this work as an experimental information retrieval framework consisting of multiple tables including various formulae which allow a plethora of syntheses of similarity functions, term weights, relative term frequencies, document weights, bidirectional relevance feedback and history adjusted term weights. The experimental results of this work reveal a better correspondence between the *ANACALYPSE* academic system ranking of documents and users' preferences than the feedback-free ranking of *Google*, a leading-edge commercial system.

- The experimental results for the various similarity scoring methods reveal that the *Cosine* and *Jaccard* similarity scoring methods produce significantly better results than the *DotProduct* and *Overlap* methods when assessing the similarity between an *ANACALYPSE Metagram* and an unseen document.

- Furthermore, the experimental results reveal a higher satisfaction amongst male users rather than female users of the *ANACALYPSE* system. Also, the similarity scoring method used and certain user characteristics, exhibit an interaction, specifically younger users seem to be more satisfied with the results of the *Cosine* method than the *Jaccard* method.

- The evaluation of history tracking of the documents visited from a root page reveals better system ranking of documents than tracking free information retrieval. Forming a *Metagram* including all the documents visited during an information retrieval session is more beneficial than forming a *Metagram* with only the root documents.

- The evaluation of stemming reveals that system information retrieval effectiveness remains unaffected. Performing stemming to extract the roots of the words and leaving the words complete produce nearly the same results in terms of information retrieval effectiveness.

- The evaluation of stop word removal reveals that this process does not appear to be beneficial and can sometimes be harmful to information retrieval system effectiveness.

7.3 Synopsis

Initially when this work was commenced one of the early aims was to find out about retrieval results compared when using plain text and when interacting with modern multi-media information retrieval systems.

It became very soon apparent that users were simply accustomed to multi-media environments and as a result they expected this pan-sensual experience as the minimum standard of service.

Hence modern users are showing plain antipathy for pure text information retrieval systems. Another interesting finding is that the standard information retrieval systems procedures involving stop words removal have a slight adverse effect on retrieval effectiveness.

However, stemming does not show any adverse effects and is mostly neutral. The idea of bidirectional fuzzy relevance feedback improves information retrieval beyond the shadow of a doubt as the experimental results show.

Furthermore the enhancement of the system knowledge with the history of web roaming by the searcher also proves beneficial for retrieval efficiency. From the aforementioned statistics it is easy to see that user gender, age, as well as system stop words removal and stemming play a role in the information retrieval process.

It appears that males are more satisfied with the results of the *ANACALYPSE* process than females. System stop-word removal, which is almost a universal norm in information retrieval, does not seem to be beneficial and can sometimes be harmful.

System stemming does not seem to be harmful or beneficial; it is simply neutral to the effectiveness of the information retrieval process. In general, while current commercial information retrieval systems provide quite adequate methods of general heuristic for information they can be limited in effectiveness and some times restrictive in query elasticity, conditions which represent additional effort in time and reading for the information seeker.

In the work of this thesis novel syntheses of bidirectional fuzzy logic and information retrieval methodologies have been developed, analyzed and evaluated by subject matter experts. These novel methodologies have been designed and implemented in an experimental information retrieval system.

According to the experimental results the academic information retrieval system *ANACALYPSE* is closer to the expert's relevance judgements compared to Google, the commercial information retrieval system. The proposed methodologies elucidate the bases for improved effectiveness and efficiency of information retrieval.

7.4 Suggestions for future work

An area of interest for future work is a distributed information retrieval system based on the same principles introduced herein and utilizing the Grid computing architecture. This strategy should alleviate two areas of concerns for modern information retrieval systems, speed of processing and retrieval and keeping always up to date current information in the database. For instance, often documents retrieved from leading commercial information retrieval systems have changed location or do not exist anymore. This distributed approach addresses this issue offering a minimal maintenance document database.

In the work of this thesis the information retrieval system learns and adapts to user behaviour. This paradigm is also useful to email filtering and categorization. For instance with a "spam" button the user discards all unsolicited email and the system learns from the contents of spam emails by building Metagrams and comparing them to new emails to identify automatically similar spam received emails in the future.

Also, further email categorization can be applied in the same manner. For instance a user may want to categorize incoming email to folders "family", "friends", "office", "personal", etc. then through a variety of Metagrams the system learns from the contents of already categorized emails to identify automatically similar received emails in the future and direct them to the appropriate folder.

Another area of interest for further work is data mining of the internet. For instance finding a specific item for sale on the Internet and building Metagrams to compare new items characteristics such as prices, specifications, terms and conditions.

Another area of interest for further work is image information retrieval. For instance in order to compare the captions and descriptions of the images with the image content, automatic recognition of the image by building Metagrams with distinguishing image characteristics may benefit information retrieval.

For instance, a more practical example of image information retrieval is to combine data mining and image retrieval to find a specific visual catalogue item for sale on the Internet and compare prices, specifications, terms and conditions.

Finally in the same area of document information retrieval a follow up to the current study would greatly benefit from a greater number of human subjects in order to gain more insight on the behaviour of a greater user population and the effect of the novel bidirectional relevance feedback of a significant number of users on information retrieval system effectiveness.

7.5 Limitations

The experimental strategy of engaging human subjects in the experimental procedure has advantages and disadvantages. The advantage in this case is that the distribution of the selected user population is both uniform and heterogeneous.

These heterogeneous and uniform characteristics allow an opportunity to study the behaviour of different user groups. However the disadvantage is that the scale of the experiments is limited to the maximum number of human subjects which is very dear to increase substantially.

Hence during the course of the experiments three different user groups were engaged of eight users each for a total of twenty four users. The first experiment had a preliminary character with eight users and a limited number of queries in an attempt to identify execution difficulties.

For this reason the results of this preliminary pilot study were not included in the work of this thesis as being too limited. However, the results of the remaining two extensive experiments are included in the Appendixes A and B under Data Tables A and B respectively. These two experiments engaged a total number of sixteen different human subjects.

Nonetheless the author would like to acknowledge that even twenty four different users is not a great number of human subjects which is the primary limitation of this user-centred approach. Further experimentation with the ability to collect data from a

scalable, much greater number of human subjects requires a new strategy for administration and execution of the experiments and remains as future work.

Appendix A – Data Table A

ID	Query	Stopwords Removal	Term Stemming	Anacalypse correlation	Google correlation	Difference AminusG	Anacalypse Change
E1	1	0	0	0.4303	-0.1878	0.6181	31%
E1	2	1	0	0.6242	-0.1878	0.812	41%
E1	3	0	1	0.4303	-0.1878	0.6181	31%
E1	4	1	1	0.6484	-0.1878	0.8362	42%
E1	5	0	0	0.2727	0.3939	-0.1212	-6%
E1	6	1	0	0.4424	0.3939	0.0485	2%
E1	7	0	1	0.4424	0.3939	0.0485	2%
E1	8	1	1	0.5272	0.3939	0.1333	7%
E1	9	0	0	0.103	0.2606	-0.1576	-8%
E1	10	1	0	-0.2121	0.2606	-0.4727	-24%
E1	11	0	1	0.103	0.2606	-0.1576	-8%
E1	12	1	1	-0.2	0.2606	-0.4606	-23%
E1	13	0	0	-0.0666	-0.0545	-0.0121	-1%
E1	14	1	0	-0.2242	-0.0545	-0.1697	-8%
E1	15	0	1	-0.0787	-0.0545	-0.0242	-1%
E1	16	1	1	-0.2606	-0.0545	-0.2061	-10%
E2	17	0	0	1	-0.0424	1.0424	52%
E2	18	1	0	0.8545	-0.0424	0.8969	45%
E2	19	0	1	0.9757	-0.0424	1.0181	51%
E2	20	1	1	0.5272	-0.0424	0.5696	28%
E2	21	0	0	1	0.1636	0.8364	42%
E2	22	1	0	0.0787	0.1636	-0.0849	-4%
E2	23	0	1	0.9151	0.1636	0.7515	38%
E2	24	1	1	0.0787	0.1636	-0.0849	-4%
E2	25	0	0	1	-0.1393	1.1393	57%
E2	26	1	0	0.6606	-0.1393	0.7999	40%
E2	27	0	1	0.9878	-0.1393	1.1271	56%
E2	28	1	1	0.6727	-0.1393	0.812	41%
E2	29	0	0	1	0.0909	0.9091	45%
E2	30	1	0	0.709	0.0909	0.6181	31%
E2	31	0	1	1	0.0909	0.9091	45%
E2	32	1	1	0.7212	0.0909	0.6303	32%
E3	33	0	0	1	0.7212	0.2788	14%
E3	34	1	0	0.8424	0.7212	0.1212	6%
E3	35	0	1	0.9636	0.7212	0.2424	12%
E3	36	1	1	0.8787	0.7212	0.1575	8%
E3	37	0	0	0.2848	-0.2606	0.5454	27%
E3	38	1	0	0.503	-0.2606	0.7636	38%
E3	39	0	1	0.2848	-0.2606	0.5454	27%
E3	40	1	1	0.4545	-0.2606	0.7151	36%
E3	41	0	0	1	0.3696	0.6304	32%
E3	42	1	0	0.5636	0.3696	0.194	10%
E3	43	0	1	0.9515	0.3696	0.5819	29%
E3	44	1	1	0.5393	0.3696	0.1697	8%
E3	45	0	0	1	0.6969	0.3031	15%
E3	46	1	0	0.709	0.6969	0.0121	1%
E3	47	0	1	0.9636	0.6969	0.2667	13%
E3	48	1	1	0.6484	0.6969	-0.0485	-2%
E4	49	0	0	0.406	0.3212	0.0848	4%
E4	50	1	0	0.3333	0.3212	0.0121	1%

ID	Query	Stopwords Removal	Term Stemming	Anacalypse correlation	Google correlation	Difference AminusG	Anacalypse Change
E4	51	0	1	0.4424	0.3212	0.1212	6%
E4	52	1	1	0.3696	0.3212	0.0484	2%
E4	53	0	0	0.6363	0.0181	0.6182	31%
E4	54	1	0	0.6969	0.0181	0.6788	34%
E4	55	0	1	0.709	0.0181	0.6909	35%
E4	56	1	1	0.7212	0.0181	0.7031	35%
E4	57	0	0	0.5515	0.406	0.1455	7%
E4	58	1	0	0.3818	0.406	-0.0242	-1%
E4	59	0	1	0.2848	0.406	-0.1212	-6%
E4	60	1	1	0.3696	0.406	-0.0364	-2%
E4	61	0	0	0.3575	0.6	-0.2425	-12%
E4	62	1	0	0.4303	0.6	-0.1697	-8%
E4	63	0	1	0.3575	0.6	-0.2425	-12%
E4	64	1	1	0.5393	0.6	-0.0607	-3%
E5	65	0	0	0.8181	0.9151	-0.097	-5%
E5	66	1	0	0.406	0.9151	-0.5091	-25%
E5	67	0	1	0.8181	0.9151	-0.097	-5%
E5	68	1	1	0.4545	0.9151	-0.4606	-23%
E5	69	0	0	-0.2242	0.2727	-0.4969	-25%
E5	70	1	0	-0.2	0.2727	-0.4727	-24%
E5	71	0	1	-0.2242	0.2727	-0.4969	-25%
E5	72	1	1	-0.2363	0.2727	-0.509	-25%
E5	73	0	0	0.2242	0.0545	0.1697	8%
E5	74	1	0	0.2848	0.0545	0.2303	12%
E5	75	0	1	0.2969	0.0545	0.2424	12%
E5	76	1	1	0.2969	0.0545	0.2424	12%
E5	77	0	0	0.1878	-0.006	0.1938	10%
E5	78	1	0	0.1272	-0.006	0.1332	7%
E5	79	0	1	0.1878	-0.006	0.1938	10%
E5	80	1	1	0.0909	-0.006	0.0969	5%
E6	81	0	0	1	-0.2242	1.2242	61%
E6	82	1	0	0.7575	-0.2242	0.9817	49%
E6	83	0	1	0.9878	-0.2242	1.212	61%
E6	84	1	1	0.806	-0.2242	1.0302	52%
E6	85	0	0	1	-0.5393	1.5393	77%
E6	86	1	0	0.5515	-0.5393	1.0908	55%
E6	87	0	1	1	-0.5393	1.5393	77%
E6	88	1	1	0.6606	-0.5393	1.1999	60%
E6	89	0	0	1	-0.1151	1.1151	56%
E6	90	1	0	0.806	-0.1151	0.9211	46%
E6	91	0	1	0.9757	-0.1151	1.0908	55%
E6	92	1	1	0.7818	-0.1151	0.8969	45%
E6	93	0	0	1	0.2606	0.7394	37%
E6	94	1	0	0.4666	0.2606	0.206	10%
E6	95	0	1	0.9515	0.2606	0.6909	35%
E6	96	1	1	0.3696	0.2606	0.109	5%
E7	97	0	0	1	-0.0424	1.0424	52%
E7	98	1	0	0.7818	-0.0424	0.8242	41%
E7	99	0	1	0.9878	-0.0424	1.0302	52%
E7	100	1	1	0.7454	-0.0424	0.7878	39%
E7	101	0	0	0.1393	0.1151	0.0242	1%
E7	102	1	0	0.2121	0.1151	0.097	5%

ID	Query	Stopwords Removal	Term Stemming	Anacalypse correlation	Google correlation	Difference AminusG	Anacalypse Change
E7	103	0	1	0.1636	0.1151	0.0485	2%
E7	104	1	1	0.1393	0.1151	0.0242	1%
E7	105	0	0	0.7212	0.4545	0.2667	13%
E7	106	1	0	0.0787	0.4545	-0.3758	-19%
E7	107	0	1	0.5393	0.4545	0.0848	4%
E7	108	1	1	0.0303	0.4545	-0.4242	-21%
E7	109	0	0	0.1636	0.3333	-0.1697	-8%
E7	110	1	0	0.0909	0.3333	-0.2424	-12%
E7	111	0	1	0.2	0.3333	-0.1333	-7%
E7	112	1	1	0.1515	0.3333	-0.1818	-9%
E8	113	0	0	0.1757	0.2363	-0.0606	-3%
E8	114	1	0	0.2606	0.2363	0.0243	1%
E8	115	0	1	0.1757	0.2363	-0.0606	-3%
E8	116	1	1	0.2121	0.2363	-0.0242	-1%
E8	117	0	0	0.2	0.4909	-0.2909	-15%
E8	118	1	0	0.0181	0.4909	-0.4728	-24%
E8	119	0	1	0.2	0.4909	-0.2909	-15%
E8	120	1	1	0.0181	0.4909	-0.4728	-24%
E8	121	0	0	-0.0424	-0.0909	0.0485	2%
E8	122	1	0	0.2727	-0.0909	0.3636	18%
E8	123	0	1	-0.006	-0.0909	0.0849	4%
E8	124	1	1	0.103	-0.0909	0.1939	10%
E8	125	0	0	0.1878	0.6121	-0.4243	-21%
E8	126	1	0	-0.0787	0.6121	-0.6908	-35%
E8	127	0	1	0.103	0.6121	-0.5091	-25%
E8	128	1	1	-0.0666	0.6121	-0.6787	-34%
Average	Values:			+0.45811	+0.190144	+0.247966	+12%

The above table presents a summary of the experimental methodology framework syntheses BAN-ABB-BBB signifying retrieval data for *12,800* documents processed.

Appendix B – Data Table B

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
1	A1	0	0	N	0	1	0.054545	0.163636	0.575758	-0.22424
1	A2	1	0	N	0	0.757576	0.054545	0.054545	0.612121	-0.41818
1	A3	0	1	N	0	0.951515	0.054545	0.151515	0.575758	-0.26061
1	A4	1	1	N	0	0.745455	0.054545	0.115152	0.527273	-0.41818
1	A5	0	0	A	12	1	-0.15152	0.418182	0.757576	0.212121
1	A6	1	0	A	12	0.769697	-0.15152	0.50303	0.854545	0.163636
1	A7	0	1	A	12	0.975758	-0.15152	0.369697	0.757576	0.175758
1	A8	1	1	A	12	0.733333	-0.15152	0.478788	0.842424	0.163636
1	A9	0	0	B	12	1	-0.15152	0.29697	0.612121	-0.05455
1	A10	1	0	B	12	0.733333	-0.15152	0.333333	0.806061	-0.01818
1	A11	0	1	B	12	0.987879	-0.15152	0.260606	0.745455	-0.11515
1	A12	1	1	B	12	0.733333	-0.15152	0.333333	0.806061	0.030303
1	A13	0	0	C	12	1	-0.07879	0.236364	0.551515	-0.12727
1	A14	1	0	C	12	0.709091	-0.07879	0.163636	0.733333	-0.17576
1	A15	0	1	C	12	0.987879	-0.07879	0.163636	0.709091	-0.15152
1	A16	1	1	C	12	0.733333	-0.07879	0.175758	0.721212	-0.0303
1	A17	0	0	D	12	1	-0.15152	0.29697	0.612121	-0.05455
1	A18	1	0	D	12	0.733333	-0.15152	0.333333	0.806061	-0.04242
1	A19	0	1	D	12	0.987879	-0.15152	0.260606	0.745455	-0.11515
1	A20	1	1	D	12	0.733333	-0.15152	0.333333	0.806061	0.006061
1	A21	0	0	E	12	1	-0.15152	0.29697	0.612121	-0.05455
1	A22	1	0	E	12	0.733333	-0.15152	0.333333	0.806061	-0.04242
1	A23	0	1	E	12	0.987879	-0.15152	0.260606	0.745455	-0.11515
1	A24	1	1	E	12	0.733333	-0.15152	0.333333	0.806061	0.006061
1	B1	0	0	N	0	1	-0.11515	0.393939	0.951515	0.684848
1	B2	1	0	N	0	0.721212	-0.11515	0.454545	0.793939	0.248485
1	B3	0	1	N	0	0.987879	-0.11515	0.466667	0.951515	0.684848
1	B4	1	1	N	0	0.6	-0.11515	0.515152	0.781818	0.345455
1	B5	0	0	A	4	1	-0.11515	0.29697	0.951515	0.527273
1	B6	1	0	A	4	0.563636	-0.11515	0.284848	0.90303	0.406061
1	B7	0	1	A	4	0.987879	-0.11515	0.406061	0.90303	0.612121
1	B8	1	1	A	4	0.490909	-0.11515	0.454545	0.866667	0.490909
1	B9	0	0	B	4	1	-0.21212	0.284848	0.939394	0.515152
1	B10	1	0	B	4	0.624242	-0.21212	0.212121	0.890909	0.29697
1	B11	0	1	B	4	1	-0.21212	0.393939	0.890909	0.6
1	B12	1	1	B	4	0.551515	-0.21212	0.406061	0.854545	0.29697
1	B13	0	0	C	4	1	-0.21212	0.284848	0.951515	0.515152
1	B14	1	0	C	4	0.624242	-0.21212	0.187879	0.890909	0.29697
1	B15	0	1	C	4	0.987879	-0.21212	0.393939	0.890909	0.6
1	B16	1	1	C	4	0.551515	-0.21212	0.442424	0.854545	0.29697
1	B17	0	0	D	4	1	-0.11515	0.29697	0.951515	0.527273
1	B18	1	0	D	4	0.284848	-0.11515	0.284848	0.90303	0.406061
1	B19	0	1	D	4	0.987879	-0.11515	0.406061	0.90303	0.612121
1	B20	1	1	D	4	0.490909	-0.11515	0.454545	0.866667	0.490909
1	B21	0	0	E	4	1	-0.11515	0.29697	0.951515	0.527273
1	B22	1	0	E	4	0.563636	-0.11515	0.284848	0.90303	0.406061
1	B23	0	1	E	4	0.987879	-0.11515	0.406061	0.90303	0.612121
1	B24	1	1	E	4	0.490909	-0.11515	0.454545	0.866667	0.490909
1	C1	0	0	N	0	1	0.430303	-0.236364	0.927273	-0.15152
1	C2	1	0	N	0	0.648485	0.430303	-0.369697	0.890909	0.163636
1	C3	0	1	N	0	0.963636	0.430303	-0.139394	0.878788	0.175758

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
1	C4	1	1	N	0	0.709091	0.430303	-0.018182	0.878788	0.175758
1	C5	0	0	A	7	1	0.430303	-0.333333	0.927273	-0.16364
1	C6	1	0	A	7	0.709091	0.430303	-0.345455	0.769697	0.248485
1	C7	0	1	A	7	0.963636	0.430303	-0.127273	0.915152	0.078788
1	C8	1	1	A	7	0.757576	0.430303	-0.030303	0.878788	0.272727
1	C9	0	0	B	7	1	0.430303	-0.357576	0.927273	-0.12727
1	C10	1	0	B	7	0.709091	0.430303	-0.406061	0.769697	0.224242
1	C11	0	1	B	7	0.224242	0.430303	-0.127273	0.915152	0.175758
1	C12	1	1	B	7	0.769697	0.430303	-0.018182	0.878788	0.272727
1	C13	0	0	C	7	1	0.430303	-0.369697	0.927273	-0.12727
1	C14	1	0	C	7	0.709091	0.430303	-0.418182	0.818182	0.248485
1	C15	0	1	C	7	0.963636	0.430303	-0.030303	0.915152	0.175758
1	C16	1	1	C	7	0.781818	0.430303	-0.018182	0.878788	0.272727
1	C17	0	0	D	7	1	0.430303	-0.357576	0.927273	-0.15152
1	C18	1	0	D	7	0.709091	0.430303	-0.29697	0.769697	0.248485
1	C19	0	1	D	7	0.963636	0.430303	-0.127273	0.915152	0.175758
1	C20	1	1	D	7	0.757576	0.430303	-0.018182	0.878788	0.272727
1	C21	0	0	E	7	1	0.430303	-0.357576	0.927273	-0.15152
1	C22	1	0	E	7	0.709091	0.430303	-0.29697	0.769697	0.248485
1	C23	0	1	E	7	0.963636	0.430303	-0.127273	0.915152	0.175758
1	C24	1	1	E	7	0.757576	0.430303	-0.018182	0.878788	0.272727
1	D1	0	0	N	0	1	0.10303	0.151515	0.393939	-0.34545
1	D2	1	0	N	0	0.248485	0.10303	0.515152	0.660606	0.551515
1	D3	0	1	N	0	0.866667	0.10303	0.175758	0.575758	-0.41818
1	D4	1	1	N	0	0.357576	0.10303	0.563636	0.648485	0.575758
1	D5	0	0	A	1	1	0.090909	0.175758	0.345455	-0.3697
1	D6	1	0	A	1	0.284848	0.090909	0.575758	0.587879	0.539394
1	D7	0	1	A	1	0.806061	0.090909	0.187879	0.442424	-0.45455
1	D8	1	1	A	1	0.369697	0.090909	0.551515	0.612121	0.563636
1	D9	0	0	B	1	1	0.090909	0.175758	0.345455	-0.3697
1	D10	1	0	B	1	0.284848	0.090909	0.575758	0.587879	0.539394
1	D11	0	1	B	1	0.806061	0.090909	0.187879	0.527273	-0.45455
1	D12	1	1	B	1	0.369697	0.090909	0.551515	0.612121	0.563636
1	D13	0	0	C	1	1	0.090909	0.175758	0.345455	-0.3697
1	D14	1	0	C	1	0.284848	0.090909	0.575758	0.587879	0.539394
1	D15	0	1	C	1	0.806061	0.090909	0.187879	0.527273	-0.45455
1	D16	1	1	C	1	0.369697	0.090909	0.575758	0.612121	0.563636
1	D17	0	0	D	1	1	0.090909	0.175758	0.345455	-0.3697
1	D18	1	0	D	1	0.284848	0.090909	0.575758	0.587879	0.539394
1	D19	0	1	D	1	0.806061	0.090909	0.187879	0.442424	-0.45455
1	D20	1	1	D	1	0.369697	0.090909	0.551515	0.612121	0.563636
1	D21	0	0	E	1	1	0.090909	0.175758	0.345455	-0.3697
1	D22	1	0	E	1	0.284848	0.090909	0.575758	0.587879	0.539394
1	D23	0	1	E	1	0.806061	0.090909	0.187879	0.442424	-0.45455
1	D24	1	1	E	1	0.369697	0.090909	0.551515	0.612121	0.563636
2	E1	0	0	N	0	1	0.236364	0.006061	0.975758	0.272727
2	E2	1	0	N	0	0.806061	0.236364	-0.030303	0.951515	0.527273
2	E3	0	1	N	0	0.987879	0.236364	0.018182	0.975758	0.139394
2	E4	1	1	N	0	0.793939	0.236364	-0.042424	0.878788	0.527273
2	E5	0	0	A	2	1	0.236364	0.006061	0.975758	0.272727
2	E6	1	0	A	2	0.806061	0.236364	-0.030303	0.927273	0.563636
2	E7	0	1	A	2	0.987879	0.236364	0.006061	0.975758	0.139394
2	E8	1	1	A	2	0.806061	0.236364	-0.042424	-0.89091	0.527273

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
2	E9	0	0	B	2	1	0.236364	0.006061	0.975758	0.272727
2	E10	1	0	B	2	0.806061	0.236364	-0.030303	0.927273	0.563636
2	E11	0	1	B	2	0.987879	0.236364	0.006061	0.951515	0.139394
2	E12	1	1	B	2	0.806061	0.236364	-0.042424	0.890909	0.527273
2	E13	0	0	C	2	1	0.236364	0.006061	0.975758	0.272727
2	E14	1	0	C	2	0.806061	0.236364	-0.030303	0.927273	0.563636
2	E15	0	1	C	2	0.987879	0.236364	0.006061	0.951515	0.139394
2	E16	1	1	C	2	0.806061	0.236364	-0.042424	0.890909	0.478788
2	E17	0	0	D	2	1	0.236364	0.006061	0.975758	0.272727
2	E18	1	0	D	2	0.806061	0.236364	-0.030303	0.927273	0.563636
2	E19	0	1	D	2	0.987879	0.236364	0.006061	0.975758	0.139394
2	E20	1	1	D	2	0.806061	0.236364	-0.042424	0.890909	0.527273
2	E21	0	0	E	2	1	0.236364	0.006061	0.975758	0.272727
2	E22	1	0	E	2	0.806061	0.236364	-0.030303	0.927273	0.563636
2	E23	0	1	E	2	0.987879	0.236364	0.006061	0.975758	0.139394
2	E24	1	1	E	2	0.806061	0.236364	-0.042424	0.890909	0.527273
2	F1	0	0	N	0	0.454545	0.309091	-0.090909	0.078788	0.066667
2	F2	1	0	N	0	0.309091	0.309091	-0.29697	0.10303	0.10303
2	F3	0	1	N	0	0.454545	0.309091	-0.078788	0.078788	0.066667
2	F4	1	1	N	0	0.309091	0.309091	-0.345455	0.10303	0.10303
2	F5	0	0	A	4	0.454545	0.406061	0.042424	0.139394	0.236364
2	F6	1	0	A	4	0.321212	0.406061	-0.272727	0.163636	0.151515
2	F7	0	1	A	4	0.442424	0.406061	0.078788	0.139394	0.272727
2	F8	1	1	A	4	0.321212	0.406061	-0.236364	0.139394	0.151515
2	F9	0	0	B	4	0.454545	0.309091	0.054545	0.078788	0.090909
2	F10	1	0	B	4	0.309091	0.309091	-0.29697	0.10303	0.10303
2	F11	0	1	B	4	0.454545	0.309091	0.078788	0.078788	0.10303
2	F12	1	1	B	4	0.309091	0.309091	-0.29697	0.078788	0.10303
2	F13	0	0	C	4	0.454545	0.309091	0.054545	0.078788	0.066667
2	F14	1	0	C	4	0.309091	0.309091	-0.29697	0.10303	0.10303
2	F15	0	1	C	4	0.454545	0.309091	0.078788	0.078788	0.066667
2	F16	1	1	C	4	0.309091	0.309091	-0.321212	0.078788	0.10303
2	F17	0	0	D	4	0.454545	0.309091	-0.006061	0.078788	0.2
2	F18	1	0	D	4	0.309091	0.309091	-0.29697	0.10303	0.10303
2	F19	0	1	D	4	0.454545	0.309091	0.054545	0.078788	0.212121
2	F20	1	1	D	4	0.309091	0.309091	-0.29697	0.078788	0.10303
2	F21	0	0	E	4	0.454545	0.309091	-0.006061	0.078788	0.2
2	F22	1	0	E	4	0.309091	0.309091	-0.29697	0.10303	0.10303
2	F23	0	1	E	4	0.454545	0.309091	0.054545	0.078788	0.212121
2	F24	1	1	E	4	0.309091	0.309091	-0.29697	0.078788	0.10303
2	G1	0	0	N	0	0.927273	0.393939	-0.042424	0.890909	-0.0303
2	G2	1	0	N	0	0.284848	0.393939	0.212121	0.90303	0.563636
2	G3	0	1	N	0	0.915152	0.393939	-0.10303	0.854545	0.090909
2	G4	1	1	N	0	0.357576	0.393939	0.018182	0.90303	0.563636
2	G5	0	0	A	2	0.927273	0.357576	0.054545	0.890909	-0.05455
2	G6	1	0	A	2	0.369697	0.357576	0.2	0.890909	0.490909
2	G7	0	1	A	2	0.927273	0.357576	-0.115152	0.866667	0.054545
2	G8	1	1	A	2	0.418182	0.357576	-0.006061	0.890909	0.418182
2	G9	0	0	B	2	0.927273	0.393939	0.042424	0.878788	-0.0303
2	G10	1	0	B	2	0.29697	0.393939	0.115152	0.90303	0.563636
2	G11	0	1	B	2	0.915152	0.393939	-0.078788	0.878788	0.090909
2	G12	1	1	B	2	0.357576	0.393939	0.042424	0.90303	0.515152
2	G13	0	0	C	2	0.927273	0.393939	0.042424	0.878788	-0.0303

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
2	G14	1	0	C	2	0.345455	0.393939	0.115152	0.90303	0.515152
2	G15	0	1	C	2	0.90303	0.393939	-0.078788	0.878788	0.090909
2	G16	1	1	C	2	0.357576	0.393939	-0.006061	0.90303	0.515152
2	G17	0	0	D	2	0.927273	0.393939	0.042424	0.878788	-0.0303
2	G18	1	0	D	2	0.29697	0.393939	0.187879	0.90303	0.563636
2	G19	0	1	D	2	0.915152	0.393939	-0.10303	0.854545	0.090909
2	G20	1	1	D	2	0.321212	0.393939	0.018182	0.90303	0.50303
2	G21	0	0	E	2	0.927273	0.393939	0.042424	0.878788	-0.0303
2	G22	1	0	E	2	0.29697	0.393939	0.187879	0.90303	0.563636
2	G23	0	1	E	2	0.915152	0.393939	-0.10303	0.854545	0.090909
2	G24	1	1	E	2	0.321212	0.393939	0.018182	0.90303	0.50303
2	H1	0	0	N	0	0.115152	-0.09091	-0.236364	0.006061	-0.05455
2	H2	1	0	N	0	-0.11515	-0.09091	-0.345455	-0.01818	-0.26061
2	H3	0	1	N	0	-0.06667	-0.09091	-0.066667	-0.0303	-0.13939
2	H4	1	1	N	0	-0.09091	-0.09091	-0.393939	-0.0303	-0.32121
2	H5	0	0	A	4	0.115152	0.10303	0.054545	0.030303	-0.17576
2	H6	1	0	A	4	0.224242	0.10303	-0.078788	-0.00606	-0.22424
2	H7	0	1	A	4	0.10303	0.10303	-0.018182	-0.09091	-0.13939
2	H8	1	1	A	4	0.321212	0.10303	0.090909	-0.07879	-0.29697
2	H9	0	0	B	4	0.115152	0.090909	0.090909	-0.07879	-0.15152
2	H10	1	0	B	4	0.284848	0.090909	-0.006061	-0.05455	-0.26061
2	H11	0	1	B	4	0.078788	0.090909	0.187879	-0.09091	-0.15152
2	H12	1	1	B	4	0.381818	0.090909	0.212121	-0.15152	-0.30909
2	H13	0	0	C	4	0.115152	0.163636	0.29697	-0.10303	-0.05455
2	H14	1	0	C	4	0.29697	0.163636	0.187879	-0.04242	-0.18788
2	H15	0	1	C	4	0.078788	0.163636	0.369697	-0.17576	-0.05455
2	H16	1	1	C	4	0.430303	0.163636	0.321212	-0.13939	-0.22424
2	H17	0	0	D	4	0.115152	0.090909	0.090909	-0.07879	-0.15152
2	H18	1	0	D	4	0.248485	0.090909	-0.066667	-0.05455	-0.26061
2	H19	0	1	D	4	0.078788	0.090909	0.139394	-0.09091	-0.15152
2	H20	1	1	D	4	0.381818	0.090909	0.10303	-0.12727	-0.30909
2	H21	0	0	E	4	0.115152	0.090909	0.090909	-0.07879	-0.15152
2	H22	1	0	E	4	0.248485	0.090909	-0.066667	-0.05455	-0.26061
2	H23	0	1	E	4	0.078788	0.090909	0.139394	-0.09091	-0.15152
2	H24	1	1	E	4	0.381818	0.090909	0.10303	-0.12727	-0.30909
3	I1	0	0	N	0	1	0.430303	-0.29697	0.830303	0.10303
3	I2	1	0	N	0	0.357576	0.430303	0.151515	0.648485	0.224242
3	I3	0	1	N	0	0.951515	0.430303	-0.333333	0.830303	0.115152
3	I4	1	1	N	0	0.272727	0.430303	0.224242	0.648485	0.321212
3	I5	0	0	A	3	1	0.454545	-0.139394	0.866667	0.284848
3	I6	1	0	A	3	0.006061	0.454545	0.29697	0.915152	0.587879
3	I7	0	1	A	3	0.951515	0.454545	-0.163636	0.866667	0.224242
3	I8	1	1	A	3	0.006061	0.454545	-0.10303	0.878788	0.624242
3	I9	0	0	B	3	1	0.454545	-0.163636	0.866667	0.284848
3	I10	1	0	B	3	0.042424	0.454545	0.309091	0.915152	0.587879
3	I11	0	1	B	3	0.987879	0.454545	-0.163636	0.866667	0.260606
3	I12	1	1	B	3	0.042424	0.454545	0.10303	0.90303	0.624242
3	I13	0	0	C	3	1	0.442424	-0.236364	0.866667	0.2
3	I14	1	0	C	3	0.018182	0.442424	0.236364	0.939394	0.587879
3	I15	0	1	C	3	0.963636	0.442424	-0.236364	0.890909	0.175758
3	I16	1	1	C	3	-0.04242	0.442424	-0.006061	0.927273	0.624242
3	I17	0	0	D	3	1	0.454545	-0.163636	0.866667	0.284848
3	I18	1	0	D	3	0.042424	0.454545	0.29697	0.915152	0.587879

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
3	I19	0	1	D	3	0.951515	0.454545	-0.187879	0.866667	0.2
3	I20	1	1	D	3	0.042424	0.454545	-0.006061	0.878788	0.624242
3	I21	0	0	E	3	1	0.454545	-0.163636	0.866667	0.284848
3	I22	1	0	E	3	0.042424	0.454545	0.29697	0.915152	0.587879
3	I23	0	1	E	3	0.951515	0.454545	-0.187879	0.866667	0.2
3	I24	1	1	E	3	0.042424	0.454545	-0.006061	0.878788	0.624242
3	J1	0	0	N	0	1	0.248485	-0.042424	1	0.187879
3	J2	1	0	N	0	0.369697	0.248485	-0.042424	1	0.442424
3	J3	0	1	N	0	1	0.248485	-0.042424	1	0.187879
3	J4	1	1	N	0	0.369697	0.248485	-0.042424	1	0.442424
3	J5	0	0	A	2	1	0.248485	-0.042424	1	0.224242
3	J6	1	0	A	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J7	0	1	A	2	1	0.248485	0.018182	1	0.224242
3	J8	1	1	A	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J9	0	0	B	2	1	0.248485	-0.042424	1	0.224242
3	J10	1	0	B	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J11	0	1	B	2	1	0.248485	0.018182	1	0.224242
3	J12	1	1	B	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J13	0	0	C	2	1	0.248485	-0.042424	1	0.224242
3	J14	1	0	C	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J15	0	1	C	2	1	0.248485	0.018182	1	0.224242
3	J16	1	1	C	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J17	0	0	D	2	1	0.248485	-0.042424	1	0.224242
3	J18	1	0	D	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J19	0	1	D	2	1	0.248485	0.018182	1	0.224242
3	J20	1	1	D	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J21	0	0	E	2	1	0.248485	-0.042424	1	0.224242
3	J22	1	0	E	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	J23	0	1	E	2	1	0.248485	0.018182	1	0.224242
3	J24	1	1	E	2	0.442424	0.248485	0.175758	0.987879	0.442424
3	K1	0	0	N	0	1	0.030303	0.030303	0.781818	0.236364
3	K2	1	0	N	0	0.575758	0.030303	0.345455	0.854545	0.2
3	K3	0	1	N	0	0.878788	0.030303	0.078788	0.830303	0.29697
3	K4	1	1	N	0	0.6	0.030303	0.272727	0.769697	0.236364
3	K5	0	0	A	1	1	0.030303	0.163636	0.951515	0.745455
3	K6	1	0	A	1	0.515152	0.030303	0.478788	0.927273	0.260606
3	K7	0	1	A	1	0.927273	0.030303	0.163636	0.890909	0.842424
3	K8	1	1	A	1	0.345455	0.030303	0.478788	0.878788	0.260606
3	K9	0	0	B	1	1	-0.01818	0.29697	0.963636	0.854545
3	K10	1	0	B	1	0.406061	-0.01818	0.442424	0.939394	0.309091
3	K11	0	1	B	1	0.939394	-0.01818	0.260606	0.866667	0.878788
3	K12	1	1	B	1	0.321212	-0.01818	0.466667	0.915152	0.309091
3	K13	0	0	C	1	1	-0.01818	0.309091	0.90303	0.866667
3	K14	1	0	C	1	0.406061	-0.01818	0.466667	0.939394	0.321212
3	K15	0	1	C	1	0.951515	-0.01818	0.333333	0.866667	0.878788
3	K16	1	1	C	1	0.321212	-0.01818	0.466667	0.890909	0.309091
3	K17	0	0	D	1	1	0.030303	0.248485	0.927273	0.745455
3	K18	1	0	D	1	0.515152	0.030303	0.454545	0.915152	0.260606
3	K19	0	1	D	1	0.951515	0.030303	0.248485	0.890909	0.842424
3	K20	1	1	D	1	0.454545	0.030303	0.478788	0.878788	0.260606
3	K21	0	0	E	1	1	0.030303	0.248485	0.927273	0.745455
3	K22	1	0	E	1	0.515152	0.030303	0.454545	0.915152	0.260606
3	K23	0	1	E	1	0.951515	0.030303	0.248485	0.890909	0.842424

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
3	K24	1	1	E	1	0.454545	0.030303	0.478788	0.878788	0.260606
3	L1	0	0	N	0	0.987879	-0.26061	-0.212121	0.890909	-0.05455
3	L2	1	0	N	0	0.830303	-0.26061	-0.490909	0.951515	-0.62424
3	L3	0	1	N	0	0.987879	-0.26061	-0.284848	0.951515	-0.00606
3	L4	1	1	N	0	0.733333	-0.26061	-0.466667	0.951515	-0.68485
3	L5	0	0	A	4	0.987879	-0.26061	-0.151515	0.890909	-0.10303
3	L6	1	0	A	4	0.709091	-0.26061	-0.466667	0.915152	-0.53939
3	L7	0	1	A	4	0.987879	-0.26061	-0.236364	0.890909	0.006061
3	L8	1	1	A	4	0.684848	-0.26061	-0.478788	0.927273	-0.40606
3	L9	0	0	B	4	0.987879	-0.26061	-0.163636	0.866667	-0.10303
3	L10	1	0	B	4	0.709091	-0.26061	-0.478788	0.915152	-0.47879
3	L11	0	1	B	4	0.987879	-0.26061	-0.054545	0.890909	0.006061
3	L12	1	1	B	4	0.709091	-0.26061	-0.381818	0.927273	-0.40606
3	L13	0	0	C	4	0.987879	-0.26061	-0.163636	0.866667	-0.0303
3	L14	1	0	C	4	0.769697	-0.26061	-0.478788	0.915152	-0.47879
3	L15	0	1	C	4	0.987879	-0.26061	-0.054545	0.890909	0.006061
3	L16	1	1	C	4	0.769697	-0.26061	-0.381818	0.927273	-0.40606
3	L17	0	0	D	4	0.987879	-0.26061	-0.151515	0.890909	-0.10303
3	L18	1	0	D	4	0.709091	-0.26061	-0.466667	0.915152	-0.53939
3	L19	0	1	D	4	0.987879	-0.26061	-0.151515	0.890909	0.006061
3	L20	1	1	D	4	0.709091	-0.26061	-0.478788	0.927273	-0.40606
3	L21	0	0	E	4	0.987879	-0.26061	-0.151515	0.890909	-0.10303
3	L22	1	0	E	4	0.709091	-0.26061	-0.466667	0.915152	-0.53939
3	L23	0	1	E	4	0.987879	-0.26061	-0.151515	0.890909	0.006061
3	L24	1	1	E	4	0.709091	-0.26061	-0.478788	0.927273	-0.40606
4	M1	0	0	N	0	0.454545	0.042424	0.527273	0.345455	0.454545
4	M2	1	0	N	0	0.042424	0.042424	0.115152	0.272727	0.10303
4	M3	0	1	N	0	0.406061	0.042424	0.369697	0.333333	0.381818
4	M4	1	1	N	0	-0.05455	0.042424	0.115152	0.272727	0.054545
4	M5	0	0	A	3	0.454545	0.042424	0.660606	0.309091	0.454545
4	M6	1	0	A	3	0.10303	0.042424	0.272727	0.212121	0.10303
4	M7	0	1	A	3	0.430303	0.042424	0.430303	0.345455	0.369697
4	M8	1	1	A	3	-0.07879	0.042424	0.2	0.272727	0.054545
4	M9	0	0	B	3	0.454545	0.054545	0.684848	0.357576	0.527273
4	M10	1	0	B	3	-0.01818	0.054545	0.430303	0.321212	0.260606
4	M11	0	1	B	3	0.406061	0.054545	0.612121	0.357576	0.515152
4	M12	1	1	B	3	-0.13939	0.054545	0.430303	0.309091	0.212121
4	M13	0	0	C	3	0.454545	0.054545	0.636364	0.272727	0.527273
4	M14	1	0	C	3	-0.01818	0.054545	0.430303	0.284848	0.260606
4	M15	0	1	C	3	0.418182	0.054545	0.612121	0.357576	0.515152
4	M16	1	1	C	3	-0.13939	0.054545	0.430303	0.309091	0.212121
4	M17	0	0	D	3	0.454545	0.090909	0.587879	0.29697	0.430303
4	M18	1	0	D	3	0.006061	0.090909	0.345455	0.248485	0.163636
4	M19	0	1	D	3	0.430303	0.090909	0.551515	0.309091	0.381818
4	M20	1	1	D	3	-0.0303	0.090909	0.272727	0.236364	0.115152
4	M21	0	0	E	3	0.454545	0.090909	0.587879	0.29697	0.430303
4	M22	1	0	E	3	0.006061	0.090909	0.345455	0.248485	0.163636
4	M23	0	1	E	3	0.430303	0.090909	0.551515	0.309091	0.381818
4	M24	1	1	E	3	-0.0303	0.090909	0.272727	0.236364	0.115152
4	N1	0	0	N	0	1	0.018182	0.684848	0.781818	0.842424
4	N2	1	0	N	0	0.624242	0.018182	0.309091	0.733333	0.2
4	N3	0	1	N	0	0.963636	0.018182	0.575758	0.781818	0.757576
4	N4	1	1	N	0	0.672727	0.018182	0.393939	0.793939	0.2

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
4	N5	0	0	A	2	1	0.127273	-0.090909	0.90303	0.672727
4	N6	1	0	A	2	0.587879	0.127273	-0.333333	0.842424	-0.0303
4	N7	0	1	A	2	0.854545	0.127273	-0.090909	0.878788	0.684848
4	N8	1	1	A	2	0.745455	0.127273	-0.212121	0.878788	-0.00606
4	N9	0	0	B	2	1	0.066667	-0.042424	0.806061	0.624242
4	N10	1	0	B	2	0.69697	0.066667	-0.29697	0.781818	-0.0303
4	N11	0	1	B	2	0.866667	0.066667	-0.006061	0.806061	0.636364
4	N12	1	1	B	2	0.6	0.066667	-0.29697	0.830303	-0.07879
4	N13	0	0	C	2	1	0.078788	-0.090909	0.781818	0.69697
4	N14	1	0	C	2	0.757576	0.078788	-0.406061	0.733333	-0.17576
4	N15	0	1	C	2	0.90303	0.078788	-0.030303	0.721212	0.672727
4	N16	1	1	C	2	0.745455	0.078788	-0.345455	0.757576	-0.21212
4	N17	0	0	D	2	1	0.030303	0.006061	0.842424	0.709091
4	N18	1	0	D	2	0.684848	0.030303	-0.369697	0.793939	-0.07879
4	N19	0	1	D	2	0.90303	0.030303	0.018182	0.842424	0.672727
4	N20	1	1	D	2	0.672727	0.030303	-0.345455	0.854545	-0.06667
4	N21	0	0	E	2	1	0.030303	0.006061	0.842424	0.709091
4	N22	1	0	E	2	0.684848	0.030303	-0.369697	0.793939	-0.07879
4	N23	0	1	E	2	0.90303	0.030303	0.018182	0.842424	0.672727
4	N24	1	1	E	2	0.672727	0.030303	-0.345455	0.854545	-0.06667
4	O1	0	0	N	0	1	0.175758	0.721212	0.915152	0.260606
4	O2	1	0	N	0	0.624242	0.175758	0.478788	0.890909	-0.49091
4	O3	0	1	N	0	0.90303	0.175758	0.6	0.878788	0.284848
4	O4	1	1	N	0	0.624242	0.175758	0.490909	0.927273	-0.49091
4	O5	0	0	A	12	1	0.175758	0.69697	0.890909	0.260606
4	O6	1	0	A	12	0.660606	0.175758	0.454545	0.915152	-0.50303
4	O7	0	1	A	12	0.915152	0.175758	0.648485	0.915152	0.284848
4	O8	1	1	A	12	0.660606	0.175758	0.50303	0.915152	-0.46667
4	O9	0	0	B	12	1	0.345455	0.684848	0.866667	0.2
4	O10	1	0	B	12	0.733333	0.345455	0.587879	0.866667	-0.6
4	O11	0	1	B	12	0.963636	0.345455	0.648485	0.842424	0.151515
4	O12	1	1	B	12	0.733333	0.345455	0.672727	0.866667	-0.6
4	O13	0	0	C	12	1	0.345455	0.648485	0.842424	0.163636
4	O14	1	0	C	12	0.733333	0.345455	0.624242	0.939394	-0.6
4	O15	0	1	C	12	0.963636	0.345455	0.648485	0.842424	0.139394
4	O16	1	1	C	12	0.733333	0.345455	0.624242	0.939394	-0.61212
4	O17	0	0	D	12	1	0.236364	0.69697	0.915152	0.260606
4	O18	1	0	D	12	0.684848	0.236364	0.539394	0.915152	-0.53939
4	O19	0	1	D	12	0.939394	0.236364	0.672727	0.890909	0.236364
4	O20	1	1	D	12	0.684848	0.236364	0.648485	0.915152	-0.52727
4	O21	0	0	E	12	1	0.236364	0.69697	0.915152	0.260606
4	O22	1	0	E	12	0.684848	0.236364	0.539394	0.915152	-0.53939
4	O23	0	1	E	12	0.939394	0.236364	0.672727	0.890909	0.236364
4	O24	1	1	E	12	0.684848	0.236364	0.648485	0.915152	-0.52727
4	P1	0	0	N	0	1	0.612121	0.212121	0.866667	-0.11515
4	P2	1	0	N	0	0.878788	0.612121	-0.236364	0.806061	-0.66061
4	P3	0	1	N	0	1	0.612121	0.284848	0.866667	-0.01818
4	P4	1	1	N	0	0.878788	0.612121	-0.29697	0.793939	-0.66061
4	P5	0	0	A	8	1	0.660606	0.10303	0.890909	-0.06667
4	P6	1	0	A	8	0.830303	0.660606	-0.006061	0.842424	-0.46667
4	P7	0	1	A	8	0.951515	0.660606	0.187879	0.878788	-0.06667
4	P8	1	1	A	8	0.830303	0.660606	0.115152	0.842424	-0.45455
4	P9	0	0	B	8	1	0.709091	-0.006061	0.915152	-0.2

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
4	P10	1	0	B	8	0.757576	0.709091	-0.066667	0.866667	-0.29697
4	P11	0	1	B	8	0.963636	0.709091	0.139394	0.915152	-0.10303
4	P12	1	1	B	8	0.757576	0.709091	0.042424	0.866667	-0.29697
4	P13	0	0	C	8	1	0.709091	-0.054545	0.890909	-0.2
4	P14	1	0	C	8	0.757576	0.709091	-0.042424	0.842424	-0.27273
4	P15	0	1	C	8	0.975758	0.709091	0.078788	0.890909	-0.05455
4	P16	1	1	C	8	0.757576	0.709091	-0.042424	0.842424	-0.35758
4	P17	0	0	D	8	1	0.660606	0.212121	0.915152	-0.07879
4	P18	1	0	D	8	0.842424	0.660606	-0.018182	0.866667	-0.45455
4	P19	0	1	D	8	1	0.660606	0.284848	0.915152	-0.07879
4	P20	1	1	D	8	0.842424	0.660606	0.066667	0.866667	-0.39394
4	P21	0	0	E	8	1	0.660606	0.212121	0.915152	-0.07879
4	P22	1	0	E	8	0.842424	0.660606	-0.018182	0.866667	-0.45455
4	P23	0	1	E	8	1	0.660606	0.284848	0.915152	-0.07879
4	P24	1	1	E	8	0.842424	0.660606	0.066667	0.866667	-0.39394
5	Q1	0	0	N	0	0.030303	-0.18788	-0.515152	-0.06667	-0.52727
5	Q2	1	0	N	0	0.212121	-0.18788	0.381818	0.381818	0.321212
5	Q3	0	1	N	0	0.454545	-0.18788	0.006061	0.381818	-0.13939
5	Q4	1	1	N	0	0.236364	-0.18788	0.430303	0.369697	0.224242
5	Q5	0	0	A	5	0.030303	-0.57576	-0.030303	-0.06667	-0.40606
5	Q6	1	0	A	5	-0.11515	-0.57576	0.163636	-0.06667	-0.24848
5	Q7	0	1	A	5	0.006061	-0.57576	-0.10303	-0.11515	-0.38182
5	Q8	1	1	A	5	-0.10303	-0.57576	0.10303	-0.12727	-0.34545
5	Q9	0	0	B	5	0.030303	-0.56364	-0.333333	-0.06667	-0.46667
5	Q10	1	0	B	5	-0.12727	-0.56364	-0.066667	-0.07879	-0.32121
5	Q11	0	1	B	5	-0.00606	-0.56364	-0.236364	-0.12727	-0.49091
5	Q12	1	1	B	5	-0.17576	-0.56364	-0.066667	-0.16364	-0.34545
5	Q13	0	0	C	5	0.030303	-0.56364	-0.381818	-0.06667	-0.46667
5	Q14	1	0	C	5	-0.10303	-0.56364	-0.2	-0.07879	-0.39394
5	Q15	0	1	C	5	-0.00606	-0.56364	-0.260606	-0.13939	-0.46667
5	Q16	1	1	C	5	-0.17576	-0.56364	-0.175758	-0.17576	-0.34545
5	Q17	0	0	D	5	0.030303	-0.57576	-0.260606	-0.06667	-0.40606
5	Q18	1	0	D	5	-0.11515	-0.57576	-0.066667	-0.06667	-0.24848
5	Q19	0	1	D	5	0.006061	-0.57576	-0.175758	-0.11515	-0.44242
5	Q20	1	1	D	5	-0.10303	-0.57576	-0.066667	-0.15152	-0.3697
5	Q21	0	0	E	5	0.030303	-0.57576	-0.260606	-0.06667	-0.40606
5	Q22	1	0	E	5	-0.11515	-0.57576	-0.066667	-0.06667	-0.24848
5	Q23	0	1	E	5	0.006061	-0.57576	-0.175758	-0.11515	-0.44242
5	Q24	1	1	E	5	-0.10303	-0.57576	-0.066667	-0.15152	-0.3697
5	R1	0	0	N	0	1	0.357576	-0.10303	0.806061	-0.30909
5	R2	1	0	N	0	0.672727	0.357576	-0.054545	0.551515	-0.12727
5	R3	0	1	N	0	0.939394	0.357576	-0.10303	0.830303	-0.35758
5	R4	1	1	N	0	0.672727	0.357576	-0.066667	0.660606	-0.09091
5	R5	0	0	A	3	1	0.357576	-0.163636	0.648485	-0.28485
5	R6	1	0	A	3	0.672727	0.357576	-0.006061	0.466667	-0.13939
5	R7	0	1	A	3	0.90303	0.357576	-0.175758	0.854545	-0.35758
5	R8	1	1	A	3	0.672727	0.357576	0.018182	0.648485	-0.09091
5	R9	0	0	B	3	1	0.478788	0.042424	0.624242	-0.24848
5	R10	1	0	B	3	0.636364	0.478788	0.090909	0.575758	-0.0303
5	R11	0	1	B	3	0.927273	0.478788	-0.078788	0.866667	-0.30909
5	R12	1	1	B	3	0.612121	0.478788	0.115152	0.745455	0.018182
5	R13	0	0	C	3	1	0.478788	0.042424	0.69697	-0.26061
5	R14	1	0	C	3	0.709091	0.478788	0.090909	0.551515	-0.0303

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
5	R15	0	1	C	3	0.939394	0.478788	-0.054545	0.878788	-0.30909
5	R16	1	1	C	3	0.612121	0.478788	0.115152	0.745455	0.018182
5	R17	0	0	D	3	1	0.478788	-0.018182	0.624242	-0.24848
5	R18	1	0	D	3	0.636364	0.478788	0.090909	0.575758	-0.0303
5	R19	0	1	D	3	0.927273	0.478788	-0.078788	0.830303	-0.30909
5	R20	1	1	D	3	0.636364	0.478788	0.115152	0.721212	0.018182
5	R21	0	0	E	3	1	0.478788	-0.018182	0.624242	-0.24848
5	R22	1	0	E	3	0.636364	0.478788	0.090909	0.575758	-0.0303
5	R23	0	1	E	3	0.927273	0.478788	-0.078788	0.830303	-0.30909
5	R24	1	1	E	3	0.636364	0.478788	0.115152	0.721212	0.018182
5	S1	0	0	N	0	0.260606	-0.15152	-0.309091	0.224242	0.272727
5	S2	1	0	N	0	-0.01818	-0.15152	-0.418182	-0.2	-0.6
5	S3	0	1	N	0	0.260606	-0.15152	-0.309091	0.224242	0.139394
5	S4	1	1	N	0	0.321212	-0.15152	-0.284848	0.10303	-0.55152
5	S5	0	0	A	1	0.260606	-0.15152	-0.333333	0.224242	0.272727
5	S6	1	0	A	1	-0.01818	-0.15152	-0.418182	-0.2	-0.6
5	S7	0	1	A	1	0.260606	-0.15152	-0.309091	0.224242	0.163636
5	S8	1	1	A	1	0.321212	-0.15152	-0.284848	0.10303	-0.55152
5	S9	0	0	B	1	0.260606	-0.15152	-0.333333	0.224242	0.272727
5	S10	1	0	B	1	-0.01818	-0.15152	-0.418182	-0.2	-0.6
5	S11	0	1	B	1	0.260606	-0.15152	-0.309091	0.224242	0.163636
5	S12	1	1	B	1	0.272727	-0.15152	-0.284848	0.10303	-0.55152
5	S13	0	0	C	1	0.260606	-0.15152	-0.333333	0.224242	0.272727
5	S14	1	0	C	1	-0.01818	-0.15152	-0.418182	-0.2	-0.6
5	S15	0	1	C	1	0.260606	-0.15152	-0.309091	0.224242	0.163636
5	S16	1	1	C	1	0.272727	-0.15152	-0.284848	0.10303	-0.55152
5	S17	0	0	D	1	0.260606	-0.15152	-0.333333	0.224242	0.272727
5	S18	1	0	D	1	-0.01818	-0.15152	-0.418182	-0.2	-0.6
5	S19	0	1	D	1	0.260606	-0.15152	-0.309091	0.224242	0.163636
5	S20	1	1	D	1	0.272727	-0.15152	-0.284848	0.10303	-0.55152
5	S21	0	0	E	1	0.260606	-0.15152	-0.333333	0.224242	0.272727
5	S22	1	0	E	1	-0.01818	-0.15152	-0.418182	-0.2	-0.6
5	S23	0	1	E	1	0.260606	-0.15152	-0.309091	0.224242	0.163636
5	S24	1	1	E	1	0.272727	-0.15152	-0.284848	0.10303	-0.55152
5	T1	0	0	N	0	0.563636	0.660606	-0.236364	-0.04242	-0.3697
5	T2	1	0	N	0	0.6	0.660606	0.090909	0.042424	-0.39394
5	T3	0	1	N	0	0.515152	0.660606	-0.224242	0.042424	-0.35758
5	T4	1	1	N	0	0.393939	0.660606	0.2	0.066667	-0.40606
5	T5	0	0	A	5	0.563636	-0.15152	0.10303	0.854545	-0.0303
5	T6	1	0	A	5	0.212121	-0.15152	0.236364	0.90303	0.175758
5	T7	0	1	A	5	0.515152	-0.15152	0.018182	0.769697	-0.01818
5	T8	1	1	A	5	0.224242	-0.15152	-0.030303	0.890909	0.175758
5	T9	0	0	B	5	0.563636	-0.15152	0.10303	0.854545	-0.0303
5	T10	1	0	B	5	0.212121	-0.15152	0.2	0.90303	0.175758
5	T11	0	1	B	5	0.515152	-0.15152	0.018182	0.769697	-0.01818
5	T12	1	1	B	5	0.333333	-0.15152	0.078788	0.890909	0.175758
5	T13	0	0	C	5	0.563636	-0.15152	0.10303	0.854545	-0.0303
5	T14	1	0	C	5	0.212121	-0.15152	0.2	0.90303	0.175758
5	T15	0	1	C	5	0.515152	-0.15152	0.018182	0.769697	-0.01818
5	T16	1	1	C	5	0.333333	-0.15152	0.078788	0.890909	0.175758
5	T17	0	0	D	5	0.563636	-0.15152	0.10303	0.854545	-0.0303
5	T18	1	0	D	5	0.212121	-0.15152	0.236364	0.90303	0.175758
5	T19	0	1	D	5	0.515152	-0.15152	0.018182	0.769697	-0.01818

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
5	T20	1	1	D	5	0.224242	-0.15152	0.078788	0.890909	0.175758
5	T21	0	0	E	5	0.563636	-0.15152	0.10303	0.854545	-0.0303
5	T22	1	0	E	5	0.212121	-0.15152	0.236364	0.90303	0.175758
5	T23	0	1	E	5	0.515152	-0.15152	0.018182	0.769697	-0.01818
5	T24	1	1	E	5	0.224242	-0.15152	0.078788	0.890909	0.175758
6	U1	0	0	N	0	1	-0.21212	-0.127273	0.951515	0.10303
6	U2	1	0	N	0	0.321212	-0.21212	0.115152	0.757576	-0.07879
6	U3	0	1	N	0	0.709091	-0.21212	-0.151515	0.939394	0.078788
6	U4	1	1	N	0	0.284848	-0.21212	0.066667	0.684848	0.006061
6	U5	0	0	A	2	1	-0.23636	-0.078788	0.90303	0.066667
6	U6	1	0	A	2	0.284848	-0.23636	0.090909	0.733333	-0.05455
6	U7	0	1	A	2	0.866667	-0.23636	-0.090909	0.927273	0.054545
6	U8	1	1	A	2	0.272727	-0.23636	-0.006061	0.624242	-0.05455
6	U9	0	0	B	2	1	-0.23636	-0.078788	0.90303	0.066667
6	U10	1	0	B	2	0.284848	-0.23636	0.090909	0.733333	-0.05455
6	U11	0	1	B	2	0.866667	-0.23636	-0.090909	0.927273	0.054545
6	U12	1	1	B	2	0.272727	-0.23636	-0.006061	0.624242	-0.05455
6	U13	0	0	C	2	1	-0.23636	-0.078788	0.90303	0.090909
6	U14	1	0	C	2	0.284848	-0.23636	0.090909	0.733333	-0.05455
6	U15	0	1	C	2	0.866667	-0.23636	-0.090909	0.927273	0.054545
6	U16	1	1	C	2	0.272727	-0.23636	-0.006061	0.624242	-0.05455
6	U17	0	0	D	2	1	-0.23636	-0.078788	0.90303	0.066667
6	U18	1	0	D	2	0.284848	-0.23636	0.090909	0.733333	-0.05455
6	U19	0	1	D	2	0.866667	-0.23636	-0.090909	0.927273	0.054545
6	U20	1	1	D	2	0.272727	-0.23636	-0.006061	0.624242	-0.05455
6	U21	0	0	E	2	1	-0.23636	-0.078788	0.90303	0.066667
6	U22	1	0	E	2	0.284848	-0.23636	0.090909	0.733333	-0.05455
6	U23	0	1	E	2	0.866667	-0.23636	-0.090909	0.927273	0.054545
6	U24	1	1	E	2	0.272727	-0.23636	-0.006061	0.624242	-0.05455
6	V1	0	0	N	0	1	-0.17576	-0.212121	0.854545	-0.06667
6	V2	1	0	N	0	0.69697	-0.17576	0.006061	0.781818	0.163636
6	V3	0	1	N	0	0.987879	-0.17576	-0.248485	0.854545	-0.0303
6	V4	1	1	N	0	0.793939	-0.17576	0.006061	0.769697	0.163636
6	V5	0	0	A	2	1	-0.32121	-0.284848	0.915152	-0.28485
6	V6	1	0	A	2	0.830303	-0.32121	0.042424	0.793939	0.10303
6	V7	0	1	A	2	0.927273	-0.32121	-0.284848	0.842424	0.042424
6	V8	1	1	A	2	0.939394	-0.32121	0.042424	0.757576	0.10303
6	V9	0	0	B	2	1	-0.32121	-0.236364	0.890909	-0.28485
6	V10	1	0	B	2	0.806061	-0.32121	0.042424	0.842424	0.10303
6	V11	0	1	B	2	0.939394	-0.32121	-0.284848	0.842424	0.042424
6	V12	1	1	B	2	0.939394	-0.32121	0.10303	0.757576	0.10303
6	V13	0	0	C	2	1	-0.32121	-0.236364	0.90303	-0.28485
6	V14	1	0	C	2	0.842424	-0.32121	0.042424	0.842424	0.10303
6	V15	0	1	C	2	0.939394	-0.32121	-0.284848	0.842424	0.042424
6	V16	1	1	C	2	0.939394	-0.32121	0.10303	0.757576	0.10303
6	V17	0	0	D	2	1	-0.32121	-0.236364	0.915152	-0.28485
6	V18	1	0	D	2	0.830303	-0.32121	0.042424	0.842424	0.10303
6	V19	0	1	D	2	0.939394	-0.32121	-0.284848	0.842424	0.042424
6	V20	1	1	D	2	0.939394	-0.32121	0.042424	0.757576	0.10303
6	V21	0	0	E	2	1	-0.32121	-0.236364	0.915152	-0.28485
6	V22	1	0	E	2	0.830303	-0.32121	0.042424	0.842424	0.10303
6	V23	0	1	E	2	0.939394	-0.32121	-0.284848	0.842424	0.042424
6	V24	1	1	E	2	0.939394	-0.32121	0.042424	0.757576	0.10303

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
6	W1	0	0	N	0	1	0.369697	-0.709091	0.842424	-0.68485
6	W2	1	0	N	0	-0.09091	0.369697	-0.733333	0.757576	0.030303
6	W3	0	1	N	0	0.963636	0.369697	-0.709091	0.842424	-0.68485
6	W4	1	1	N	0	-0.21212	0.369697	-0.842424	0.806061	0.066667
6	W5	0	0	A	1	1	0.369697	-0.709091	0.842424	-0.68485
6	W6	1	0	A	1	-0.09091	0.369697	-0.733333	0.757576	0.030303
6	W7	0	1	A	1	0.963636	0.369697	-0.709091	0.842424	-0.68485
6	W8	1	1	A	1	-0.21212	0.369697	-0.842424	0.806061	0.066667
6	W9	0	0	B	1	1	0.369697	-0.709091	0.842424	-0.68485
6	W10	1	0	B	1	-0.09091	0.369697	-0.733333	0.757576	0.030303
6	W11	0	1	B	1	0.951515	0.369697	-0.709091	0.842424	-0.68485
6	W12	1	1	B	1	-0.21212	0.369697	-0.842424	0.806061	0.066667
6	W13	0	0	C	1	1	0.369697	-0.709091	0.842424	-0.68485
6	W14	1	0	C	1	-0.09091	0.369697	-0.733333	0.757576	0.030303
6	W15	0	1	C	1	0.951515	0.369697	-0.709091	0.842424	-0.68485
6	W16	1	1	C	1	-0.21212	0.369697	-0.842424	0.806061	0.066667
6	W17	0	0	D	1	1	0.369697	-0.709091	0.842424	-0.68485
6	W18	1	0	D	1	-0.09091	0.369697	-0.733333	0.757576	0.030303
6	W19	0	1	D	1	0.951515	0.369697	-0.709091	0.842424	-0.68485
6	W20	1	1	D	1	-0.21212	0.369697	-0.842424	0.806061	0.066667
6	W21	0	0	E	1	1	0.369697	-0.709091	0.842424	-0.68485
6	W22	1	0	E	1	-0.09091	0.369697	-0.733333	0.757576	0.030303
6	W23	0	1	E	1	0.951515	0.369697	-0.709091	0.842424	-0.68485
6	W24	1	1	E	1	-0.21212	0.369697	-0.842424	0.806061	0.066667
6	X1	0	0	N	0	0.018182	0.212121	0.745455	-0.04242	0.430303
6	X2	1	0	N	0	0.636364	0.212121	0.648485	-0.10303	0.272727
6	X3	0	1	N	0	0.006061	0.212121	0.709091	-0.07879	0.272727
6	X4	1	1	N	0	0.563636	0.212121	0.709091	-0.05455	0.272727
6	X5	0	0	A	2	0.018182	0.381818	0.709091	0.030303	0.660606
6	X6	1	0	A	2	0.684848	0.381818	0.684848	0.442424	0.224242
6	X7	0	1	A	2	-0.00606	0.381818	0.672727	0.030303	0.466667
6	X8	1	1	A	2	0.648485	0.381818	0.672727	0.454545	0.224242
6	X9	0	0	B	2	0.018182	0.357576	0.69697	0.006061	0.733333
6	X10	1	0	B	2	0.684848	0.357576	0.648485	0.454545	0.29697
6	X11	0	1	B	2	-0.01818	0.357576	0.684848	0.018182	0.515152
6	X12	1	1	B	2	0.587879	0.357576	0.684848	0.442424	0.29697
6	X13	0	0	C	2	0.018182	-0.00606	0.612121	0.006061	0.2
6	X14	1	0	C	2	0.151515	-0.00606	0.6	-0.10303	0.2
6	X15	0	1	C	2	0.260606	-0.00606	0.648485	0.066667	-0.09091
6	X16	1	1	C	2	0.163636	-0.00606	0.648485	-0.11515	0.2
6	X17	0	0	D	2	0.018182	0.381818	0.709091	0.030303	0.660606
6	X18	1	0	D	2	0.745455	0.381818	0.672727	0.442424	0.224242
6	X19	0	1	D	2	-0.00606	0.381818	0.69697	0.030303	0.50303
6	X20	1	1	D	2	0.648485	0.381818	0.672727	0.454545	0.224242
6	X21	0	0	E	2	0.018182	0.381818	0.709091	0.030303	0.660606
6	X22	1	0	E	2	0.745455	0.381818	0.672727	0.442424	0.224242
6	X23	0	1	E	2	-0.00606	0.381818	0.69697	0.030303	0.50303
6	X24	1	1	E	2	0.648485	0.381818	0.672727	0.454545	0.224242
7	Y1	0	0	N	0	1	-0.69697	0.309091	0.721212	0.030303
7	Y2	1	0	N	0	0.866667	-0.69697	0.406061	0.769697	0.139394
7	Y3	0	1	N	0	0.951515	-0.69697	0.151515	0.709091	-0.00606
7	Y4	1	1	N	0	0.927273	-0.69697	0.418182	0.769697	0.139394
7	Y5	0	0	A	1	1	-0.73333	0.393939	0.793939	0.139394

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
7	Y6	1	0	A	1	0.854545	-0.73333	0.466667	0.830303	0.248485
7	Y7	0	1	A	1	0.987879	-0.73333	0.345455	0.793939	0.187879
7	Y8	1	1	A	1	0.830303	-0.73333	0.442424	0.830303	0.224242
7	Y9	0	0	B	1	1	-0.73333	0.393939	0.793939	0.139394
7	Y10	1	0	B	1	0.854545	-0.73333	0.466667	0.830303	0.248485
7	Y11	0	1	B	1	0.987879	-0.73333	0.345455	0.793939	0.187879
7	Y12	1	1	B	1	0.830303	-0.73333	0.442424	0.830303	0.224242
7	Y13	0	0	C	1	1	-0.73333	0.393939	0.793939	0.139394
7	Y14	1	0	C	1	0.854545	-0.73333	0.466667	0.830303	0.248485
7	Y15	0	1	C	1	0.987879	-0.73333	0.345455	0.793939	0.187879
7	Y16	1	1	C	1	0.830303	-0.73333	0.442424	0.830303	0.224242
7	Y17	0	0	D	1	1	-0.73333	0.393939	0.793939	0.139394
7	Y18	1	0	D	1	0.854545	-0.73333	0.466667	0.830303	0.248485
7	Y19	0	1	D	1	0.987879	-0.73333	0.345455	0.793939	0.187879
7	Y20	1	1	D	1	0.830303	-0.73333	0.442424	0.830303	0.224242
7	Y21	0	0	E	1	1	-0.73333	0.393939	0.793939	0.139394
7	Y22	1	0	E	1	0.854545	-0.73333	0.466667	0.830303	0.248485
7	Y23	0	1	E	1	0.987879	-0.73333	0.345455	0.793939	0.187879
7	Y24	1	1	E	1	0.830303	-0.73333	0.442424	0.830303	0.224242
7	Z1	0	0	N	0	0.454545	0.042424	0.490909	0.478788	0.636364
7	Z2	1	0	N	0	0.090909	0.042424	0.539394	0.551515	0.672727
7	Z3	0	1	N	0	0.442424	0.042424	0.563636	0.454545	0.684848
7	Z4	1	1	N	0	0.090909	0.042424	0.648485	0.551515	0.733333
7	Z5	0	0	A	1	0.454545	0.054545	0.393939	0.515152	0.563636
7	Z6	1	0	A	1	0.10303	0.054545	0.442424	0.527273	0.6
7	Z7	0	1	A	1	0.442424	0.054545	0.393939	0.50303	0.624242
7	Z8	1	1	A	1	0.163636	0.054545	0.563636	0.527273	0.672727
7	Z9	0	0	B	1	0.454545	0.054545	0.393939	0.515152	0.563636
7	Z10	1	0	B	1	0.10303	0.054545	0.442424	0.527273	0.6
7	Z11	0	1	B	1	0.442424	0.054545	0.393939	0.50303	0.624242
7	Z12	1	1	B	1	0.163636	0.054545	0.563636	0.527273	0.672727
7	Z13	0	0	C	1	0.454545	0.054545	0.393939	0.515152	0.563636
7	Z14	1	0	C	1	0.10303	0.054545	0.442424	0.527273	0.6
7	Z15	0	1	C	1	0.442424	0.054545	0.393939	0.50303	0.624242
7	Z16	1	1	C	1	0.163636	0.054545	0.563636	0.527273	0.672727
7	Z17	0	0	D	1	0.454545	0.054545	0.393939	0.515152	0.563636
7	Z18	1	0	D	1	0.10303	0.054545	0.442424	0.527273	0.6
7	Z19	0	1	D	1	0.442424	0.054545	0.393939	0.50303	0.624242
7	Z20	1	1	D	1	0.163636	0.054545	0.563636	0.527273	0.672727
7	Z21	0	0	E	1	0.454545	0.054545	0.393939	0.515152	0.563636
7	Z22	1	0	E	1	0.10303	0.054545	0.442424	0.527273	0.6
7	Z23	0	1	E	1	0.442424	0.054545	0.393939	0.50303	0.624242
7	Z24	1	1	E	1	0.163636	0.054545	0.563636	0.527273	0.672727
7	AA1	0	0	N	0	0.915152	-0.38182	0.866667	0.915152	0.866667
7	AA2	1	0	N	0	0.975758	-0.38182	0.90303	0.927273	0.6
7	AA3	0	1	N	0	0.890909	-0.38182	0.915152	0.915152	0.927273
7	AA4	1	1	N	0	0.975758	-0.38182	0.90303	0.939394	0.6
7	AA5	0	0	A	2	0.915152	-0.39394	0.709091	0.830303	0.636364
7	AA6	1	0	A	2	0.709091	-0.39394	0.636364	0.721212	0.527273
7	AA7	0	1	A	2	0.866667	-0.39394	0.69697	0.818182	0.672727
7	AA8	1	1	A	2	0.733333	-0.39394	0.660606	0.781818	0.587879
7	AA9	0	0	B	2	0.915152	-0.32121	0.709091	0.660606	0.478788
7	AA10	1	0	B	2	0.721212	-0.32121	0.660606	0.50303	0.6

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
7	AA11	0	1	B	2	0.866667	-0.32121	0.684848	0.721212	0.490909
7	AA12	1	1	B	2	0.721212	-0.32121	0.6	0.69697	0.430303
7	AA13	0	0	C	2	0.915152	-0.32121	0.709091	0.563636	0.454545
7	AA14	1	0	C	2	0.563636	-0.32121	0.660606	0.50303	0.490909
7	AA15	0	1	C	2	0.866667	-0.32121	0.684848	0.612121	0.490909
7	AA16	1	1	C	2	0.636364	-0.32121	0.660606	0.6	0.430303
7	AA17	0	0	D	2	0.915152	-0.39394	0.709091	0.769697	0.551515
7	AA18	1	0	D	2	0.709091	-0.39394	0.551515	0.6	0.527273
7	AA19	0	1	D	2	0.866667	-0.39394	0.69697	0.757576	0.648485
7	AA20	1	1	D	2	0.709091	-0.39394	0.636364	0.624242	0.454545
7	AA21	0	0	E	2	0.915152	-0.39394	0.709091	0.769697	0.551515
7	AA22	1	0	E	2	0.709091	-0.39394	0.551515	0.6	0.527273
7	AA23	0	1	E	2	0.866667	-0.39394	0.69697	0.757576	0.648485
7	AA24	1	1	E	2	0.709091	-0.39394	0.636364	0.624242	0.454545
7	AB1	0	0	N	0	1	-0.2	-0.587879	0.854545	-0.70909
7	AB2	1	0	N	0	0.721212	-0.2	-0.490909	0.842424	0.345455
7	AB3	0	1	N	0	1	-0.2	-0.539394	0.866667	-0.6
7	AB4	1	1	N	0	0.684848	-0.2	-0.515152	0.830303	0.345455
7	AB5	0	0	A	2	1	-0.2	-0.454545	0.854545	-0.63636
7	AB6	1	0	A	2	0.757576	-0.2	-0.357576	0.781818	0.260606
7	AB7	0	1	A	2	1	-0.2	-0.406061	0.866667	-0.61212
7	AB8	1	1	A	2	0.69697	-0.2	-0.393939	0.781818	0.345455
7	AB9	0	0	B	2	1	-0.2	-0.466667	0.866667	-0.68485
7	AB10	1	0	B	2	0.793939	-0.2	-0.248485	0.781818	0.260606
7	AB11	0	1	B	2	1	-0.2	-0.406061	0.866667	-0.61212
7	AB12	1	1	B	2	0.745455	-0.2	-0.187879	0.781818	0.345455
7	AB13	0	0	C	2	1	-0.2	-0.466667	0.866667	-0.68485
7	AB14	1	0	C	2	0.793939	-0.2	-0.248485	0.781818	0.260606
7	AB15	0	1	C	2	1	-0.2	-0.333333	0.866667	-0.61212
7	AB16	1	1	C	2	0.745455	-0.2	-0.115152	0.781818	0.345455
7	AB17	0	0	D	2	1	-0.2	-0.466667	0.866667	-0.68485
7	AB18	1	0	D	2	0.745455	-0.2	-0.333333	0.781818	0.260606
7	AB19	0	1	D	2	1	-0.2	-0.406061	0.866667	-0.61212
7	AB20	1	1	D	2	0.745455	-0.2	-0.29697	0.781818	0.345455
7	AB21	0	0	E	2	1	-0.2	-0.466667	0.866667	-0.68485
7	AB22	1	0	E	2	0.745455	-0.2	-0.333333	0.781818	0.260606
7	AB23	0	1	E	2	1	-0.2	-0.406061	0.866667	-0.61212
7	AB24	1	1	E	2	0.745455	-0.2	-0.29697	0.781818	0.345455
8	AC1	0	0	N	0	0.963636	0.490909	0.660606	0.684848	0.175758
8	AC2	1	0	N	0	0.69697	0.490909	0.6	0.878788	0.69697
8	AC3	0	1	N	0	0.963636	0.490909	0.6	0.721212	0.042424
8	AC4	1	1	N	0	0.672727	0.490909	0.684848	0.878788	0.636364
8	AC5	0	0	A	1	0.963636	0.321212	0.769697	0.757576	0.236364
8	AC6	1	0	A	1	0.745455	0.321212	0.733333	0.90303	0.781818
8	AC7	0	1	A	1	0.951515	0.321212	0.733333	0.830303	0.212121
8	AC8	1	1	A	1	0.818182	0.321212	0.69697	0.90303	0.69697
8	AC9	0	0	B	1	0.963636	0.321212	0.781818	0.890909	0.709091
8	AC10	1	0	B	1	0.733333	0.321212	0.721212	0.90303	0.769697
8	AC11	0	1	B	1	0.963636	0.321212	0.781818	0.854545	0.6
8	AC12	1	1	B	1	0.806061	0.321212	0.69697	0.90303	0.721212
8	AC13	0	0	C	1	0.963636	0.321212	0.781818	0.90303	0.709091
8	AC14	1	0	C	1	0.733333	0.321212	0.721212	0.90303	0.769697
8	AC15	0	1	C	1	0.963636	0.321212	0.769697	0.890909	0.709091

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
8	AC16	1	1	C	1	0.806061	0.321212	0.721212	0.90303	0.769697
8	AC17	0	0	D	1	0.963636	0.321212	0.769697	0.818182	0.624242
8	AC18	1	0	D	1	0.745455	0.321212	0.733333	0.90303	0.769697
8	AC19	0	1	D	1	0.951515	0.321212	0.769697	0.830303	0.369697
8	AC20	1	1	D	1	0.806061	0.321212	0.69697	0.90303	0.69697
8	AC21	0	0	E	1	0.963636	0.321212	0.769697	0.818182	0.624242
8	AC22	1	0	E	1	0.745455	0.321212	0.733333	0.90303	0.769697
8	AC23	0	1	E	1	0.951515	0.321212	0.769697	0.830303	0.369697
8	AC24	1	1	E	1	0.806061	0.321212	0.69697	0.90303	0.69697
8	AD1	0	0	N	0	0.927273	0.066667	0.321212	0.636364	-0.33333
8	AD2	1	0	N	0	0.806061	0.066667	0.345455	0.709091	0.187879
8	AD3	0	1	N	0	0.878788	0.066667	0.248485	0.672727	-0.33333
8	AD4	1	1	N	0	0.818182	0.066667	0.515152	0.709091	0.175758
8	AD5	0	0	A	2	0.927273	0.018182	0.284848	0.890909	-0.0303
8	AD6	1	0	A	2	0.90303	0.018182	0.430303	0.939394	0.321212
8	AD7	0	1	A	2	0.854545	0.018182	0.406061	0.890909	-0.07879
8	AD8	1	1	A	2	0.927273	0.018182	0.430303	0.939394	0.29697
8	AD9	0	0	B	2	0.927273	0.115152	0.406061	0.672727	-0.32121
8	AD10	1	0	B	2	0.90303	0.115152	0.539394	0.745455	0.127273
8	AD11	0	1	B	2	0.806061	0.115152	0.418182	0.69697	-0.33333
8	AD12	1	1	B	2	0.854545	0.115152	0.6	0.745455	0.054545
8	AD13	0	0	C	2	0.927273	0.066667	0.381818	0.660606	-0.33333
8	AD14	1	0	C	2	0.866667	0.066667	0.515152	0.709091	0.10303
8	AD15	0	1	C	2	0.818182	0.066667	0.381818	0.660606	-0.34545
8	AD16	1	1	C	2	0.866667	0.066667	0.527273	0.709091	0.030303
8	AD17	0	0	D	2	0.927273	0.115152	0.284848	0.69697	-0.32121
8	AD18	1	0	D	2	0.90303	0.115152	0.539394	0.745455	0.127273
8	AD19	0	1	D	2	0.806061	0.115152	0.418182	0.69697	-0.33333
8	AD20	1	1	D	2	0.830303	0.115152	0.527273	0.745455	0.054545
8	AD21	0	0	E	2	0.927273	0.115152	0.284848	0.69697	-0.32121
8	AD22	1	0	E	2	0.90303	0.115152	0.539394	0.745455	0.127273
8	AD23	0	1	E	2	0.806061	0.115152	0.418182	0.69697	-0.33333
8	AD24	1	1	E	2	0.830303	0.115152	0.527273	0.745455	0.054545
8	AE1	0	0	N	0	0.224242	0.272727	0.878788	0.236364	0.830303
8	AE2	1	0	N	0	0.10303	0.272727	0.842424	0.260606	0.781818
8	AE3	0	1	N	0	0.078788	0.272727	0.769697	0.272727	0.709091
8	AE4	1	1	N	0	-0.05455	0.272727	0.781818	0.272727	0.721212
8	AE5	0	0	A	3	0.224242	-0.3697	-0.163636	-0.09091	-0.16364
8	AE6	1	0	A	3	-0.13939	-0.3697	0.006061	-0.11515	0.10303
8	AE7	0	1	A	3	0.272727	-0.3697	-0.248485	-0.05455	-0.30909
8	AE8	1	1	A	3	-0.13939	-0.3697	-0.2	-0.11515	-0.04242
8	AE9	0	0	B	3	0.224242	-0.34545	-0.248485	-0.18788	-0.24848
8	AE10	1	0	B	3	-0.23636	-0.34545	-0.139394	-0.13939	-0.01818
8	AE11	0	1	B	3	0.212121	-0.34545	-0.333333	-0.04242	-0.33333
8	AE12	1	1	B	3	-0.12727	-0.34545	-0.284848	-0.12727	-0.16364
8	AE13	0	0	C	3	0.224242	-0.35758	-0.260606	-0.17576	-0.32121
8	AE14	1	0	C	3	-0.24848	-0.35758	-0.2	-0.16364	-0.04242
8	AE15	0	1	C	3	0.2	-0.35758	-0.381818	-0.04242	-0.34545
8	AE16	1	1	C	3	-0.2	-0.35758	-0.345455	-0.16364	-0.18788
8	AE17	0	0	D	3	0.224242	-0.34545	-0.248485	-0.09091	-0.24848
8	AE18	1	0	D	3	-0.23636	-0.34545	-0.139394	-0.13939	-0.01818
8	AE19	0	1	D	3	0.284848	-0.34545	-0.333333	-0.09091	-0.33333
8	AE20	1	1	D	3	-0.11515	-0.34545	-0.284848	-0.13939	-0.16364

ID	Q	R	S	H	HF	Cosine	Google	DotProduct	Jaccard	Overlap
8	AE21	0	0	E	3	0.224242	-0.34545	-0.248485	-0.09091	-0.24848
8	AE22	1	0	E	3	-0.23636	-0.34545	-0.139394	-0.13939	-0.01818
8	AE23	0	1	E	3	0.284848	-0.34545	-0.333333	-0.09091	-0.33333
8	AE24	1	1	E	3	-0.11515	-0.34545	-0.284848	-0.13939	-0.16364
8	AF1	0	0	N	0	0.309091	0.333333	-0.212121	-0.30909	0.006061
8	AF2	1	0	N	0	0.248485	0.333333	0.018182	-0.30909	0.10303
8	AF3	0	1	N	0	0.309091	0.333333	-0.212121	-0.30909	0.006061
8	AF4	1	1	N	0	0.248485	0.333333	0.018182	-0.30909	-0.00606
8	AF5	0	0	A	1	0.309091	0.539394	-0.139394	-0.30909	-0.34545
8	AF6	1	0	A	1	0.284848	0.539394	-0.139394	-0.07879	-0.15152
8	AF7	0	1	A	1	0.309091	0.539394	-0.151515	-0.30909	-0.34545
8	AF8	1	1	A	1	0.309091	0.539394	-0.139394	-0.07879	-0.33333
8	AF9	0	0	B	1	0.309091	0.539394	-0.139394	-0.30909	-0.34545
8	AF10	1	0	B	1	0.284848	0.539394	-0.139394	-0.07879	-0.15152
8	AF11	0	1	B	1	0.309091	0.539394	-0.151515	-0.30909	-0.34545
8	AF12	1	1	B	1	0.309091	0.539394	-0.139394	-0.07879	-0.33333
8	AF13	0	0	C	1	0.309091	0.539394	-0.139394	-0.30909	-0.34545
8	AF14	1	0	C	1	0.284848	0.539394	-0.139394	-0.07879	-0.15152
8	AF15	0	1	C	1	0.309091	0.539394	-0.151515	-0.30909	-0.34545
8	AF16	1	1	C	1	0.309091	0.539394	-0.139394	-0.07879	-0.33333
8	AF17	0	0	D	1	0.309091	0.539394	-0.139394	-0.30909	-0.34545
8	AF18	1	0	D	1	0.284848	0.539394	-0.139394	-0.07879	-0.15152
8	AF19	0	1	D	1	0.309091	0.539394	-0.151515	-0.30909	-0.34545
8	AF20	1	1	D	1	0.309091	0.539394	-0.139394	-0.07879	-0.33333
8	AF21	0	0	E	1	0.309091	0.539394	-0.139394	-0.30909	-0.34545
8	AF22	1	0	E	1	0.284848	0.539394	-0.139394	-0.07879	-0.15152
8	AF23	0	1	E	1	0.309091	0.539394	-0.151515	-0.30909	-0.34545
8	AF24	1	1	E	1	0.309091	0.539394	-0.139394	-0.07879	-0.33333
Average		Values:				0.603835	0.066351	0.106692	0.611111	0.09929

The above table presents a summary of the experimental methodology framework syntheses [ABFGH]A[NABCDE]-ABB-BBB signifying retrieval data for *76,800* documents processed. In the above table ID is the user identification, Q is the query, R is the removal of stop words, S is the stemming, H is the history formula (see Chapter 4, Table 4.6) and HF is the number of history files.

Appendix C – Stop words

A
a
about
above
across
after
afterwards
again
against
all
almost
alone
along
already
also
although
always
among
amongst
an
and
another
any
anyhow
anyone
anything
anywhere
are
around
as
at

B
back
be
became
because
become
becomes
becoming
been
before
below
beside
besides
between
beyond
both
but
by

C, D, E
can
cannot
can't
co
could
couldn't
did
do
don't
down
during
each
eg
eight
either
else
elsewhere
enough
etc
even
ever
every
everyone
everything
everywhere
except

F
few
fify
first
five
for
former
formerly
forty
found
four
from
further

G, H
get
go
had
has
hasn't
have
he
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
him
himself
his
how
however
hundred

I, L
i
ie
if
in
inc
indeed
into
is
it
its
itself
last
latter
latterly
least
less
like
ltd

M
made
many
me
meanwhile
might
more
moreover
most
mostly
move
much
must
my
myself

N
namely
neither
never
nevertheless
new
next
nine
no
nobody
none
noone
nor
not
nothing
now
nowhere

O
of
off
often
on
once
one
only
onto
or
other
others
otherwise
our
ours
ourselves
out
over
own

P, R, S
part
per
perhaps
rather
same
see
seem
seemed
seeming
seems
several
she
should
since
six
sixty
so
some
somehow
someone
something
sometime
sometimes
somewhere
still
such

T
ten
than
that
the
their
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
they
this
those
though
three
through
throughout
thru
thus
to
together
too
toward
towards
twenty
two

U, V
under
until
up
upon
us
very
via

W
was
we
well
were
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
with
within
without
would

Y
yes
yet
you
your
yours
yourself
yourselves

The stop words were selected from three different sources, firstly from: *A domain based approach to natural language modelling,* PhD thesis, Queen's University of Belfast (Donnelly, 1998) secondly from the Glascow information retrieval research

group, http://www.dcs.gla.ac.uk/idom/ and finally from the Cobuild stop-word list http://www.cobuild.collins.co.uk/. The rationale behind the selection of the words is that if a term occurred in two or more of the aforementioned lists then it was included in this table.

Appendix D – Porter's stemming algorithm

The following paragraphs present an algorithm for suffix stripping which is used in the work of this thesis, is well known and very popular in the academic community due to its simplicity and good performance in terms of processing time, precision and recall (Porter, 1997).

To present the suffix stripping algorithm in its entirety we will need a few definitions. A consonant in a word is a letter other than A, E, I, O and U, and other than Y preceded by a consonant. (The fact that the term 'consonant' is defined to some extent in terms of itself does not make it ambiguous.) So in TOY the consonants are T and Y, in SYZYG Y they are S, Z and G. If a letter is not a consonant it is a vowel.

A consonant will be denoted by c, a vowel by v. A list ccc ... of length greater than 0 will be denoted by C, and a list vvv ... of length greater than 0 will be denoted by V. Any word, or part of a word, therefore has one of the four forms:

CVCV ...C
CVCV ...V
VCVC ...C
VCVC ...V

These may all be represented by the single form:

[C] VCVC ...[V]

where the square brackets denote arbitrary presence of their contents. Using $(VC)^m$ to denote VC repeated m times, this may again be written as:

[C] $(VC)^m$ [V].

Where *m* will be called the measure of any word or word part when represented in this form. The case *m*=0 covers the null word. Here are some examples:

m=0 TR, EE, TREE, Y, BY.
m=1 TROUBLE, OATS, TREES, IVY.
m=2 TROUBLES, PRIVATE, OATEN, ORRERY.

The rules for removing a suffix will be given in the form

(condition) S1 → S2

This means that if a word ends with the suffix S1, and the stem before S1 satisfies the given condition, S1 is replaced by S2. The condition is usually given in terms of m, e.g.

(m> 1) EMENT →

Here S1 is 'EMENT' and S2 is null. This would map REPLACEMENT to REPLAC, since REPLAC is word part for which m=2.

The 'condition' part may also contain the following:

*S -the stem ends with S (and similarly for the other letters).

v -the stem contains a vowel.

*d -the stem ends with a double consonant (e.g. -TT, -SS).

*o -the stem ends cvc, where the second c is not W, X or Y (e.g. -WIL, -HOP).

And the condition part may also contain expressions with and. or and not, so that

(m> 1 and (*S or *T))

tests for a stem with m> 1 ending in S or T, while

(*d and not (*L or *S or *Z))

tests for a stem ending with a double consonant other than L, S or Z. Elaborate conditions like this are required only very rarely.

In a set of rules written beneath each other, only one is obeyed, and this will be the one with the longest matching S1 for the given word.

For example, with

SSES	→	SS
IES	→	I
SS	→	SS
S	→	

(here the conditions are all null) CARESSES maps to CARESS since SSES is the longest match for S1. Equally CARESS maps to CARESS (S1 ='SS') and CARES to CARE (S1 ='S').

In the rules below, examples of their application, successful or otherwise, are given on the right in lower case. The algorithm now follows:

Step 1 a

SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
			ties	→	ti
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Step 1 b

(m>0) EED	→	EE	feed	→	feed
			agreed	→	agree
(*v*) ED	→		plastered	→	plaster
			bled	→	bled
(*v*) ING	→		motoring	→	motor
			sing	→	sing

If the second or third of the rules in Step 1 b is successful, the following is done:

AT	→	ATE			conflat(ed)	→	conflate
BL	→	BLE			troubl(ing)	→	trouble
IZ	→	IZE			siz(ed)	→	size
(*d	and	not	(*L or	*S or *Z))			
	→	single	letter		hopp(ing)	→	hop
					tann(ed)	→	tan
					fall(ing)	→	fall
					hiss(ing)	→	hiss
					fizz(ed)	→	fizz
(m=1	and	*o)→	E		fail(ing)	→	fail
					fil(ing)	→	file

The rule to map to a single letter causes the removal of one of the double letter pair. The -E is put back on -AT, -BL and -IZ, so that the suffixes -ATE, -BLE and -IZE can be recognised later. This E may be removed in step 4.

Step 1 c

(*v*) Y	→	I	happy	→	happi
			sky	→	sky

Step 1 deals with plurals and past participles. The subsequent steps are much more straightforward.

Step 2

(m>0)	ATIONAL	→	ATE	relational	→	relate
(m>0)	TIONAL	→	TION	conditional	→	condition
				rational	→	rational
(m>0)	ENCI	→	ENCE	valenci	→	valence
(m>0)	ANCI	→	ANCE	hesitanci	→	hesitance
(m>0)	IZER	→	IZE	digitizer	→	digitize
(m>0)	ABLI	→	ABLE	conformabli	→	conformable
(m>0)	ALLI	→	AL	radicalli	→	radical
(m>0)	ENTLI	→	ENT	differently	→	different
(m>0)	ELI	→	E	vileli	→	vile
(m>0)	OUSLI	→	OUS	analogousli	→	analogous
(m>0)	IZATION	→	IZE	vietnamization	→	vietnamize
(m>0)	ATION	→	ATE	predication	→	predicate
(m>0)	ATOR	→	ATE	operator	→	operate
(m>0)	ALISM	→	AL	feudalism	→	feudal
(m>0)	IVENESS	→	IVE	decisiveness	→	decisive
(m>0)	FULNESS	→	FUL	hopefulness	→	hopeful
(m>0)	OUSNESS	→	OUS	callousness	→	callous
(m>0)	ALITI	→	AL	formaliti	→	formal
(m>0)	IVITI	→	IVE	sensitiviti	→	sensitive
(m>0)	BILITI	→	BLE	sensibiliti	→	sensibile

The test for the string S1 can be made fast by doing a program switch on the penultimate letter of the word being tested. This gives a fairly even breakdown of the possible values of the string S1. It will be seen in fact that the S1 -strings in step 2 are presented here in the alphabetical order of their penultimate letter. Similar techniques may be applied in the other steps.

Step 3

(m>0)	ICATE	→	IC	triplicate	→	triplic
(m>0)	ATIVE	→		formative	→	form
(m>0)	ALIZE	→	AL	formalize	→	formal
(m>0)	ICITI	→	IC	electriciti	→	electric
(m>0)	ICAL	→	IC	electrical	→	electric
(m>0)	FUL	→		hopeful	→	hope
(m>0)	NESS	→		goodness	→	good

Step 4

(m>1)	AL	→	revival	→	reviv
(m>1)	ANCE	→	allowance	→	allow
(m>1)	ENCE	→	inference	→	infer
(m>1)	ER	→	airliner	→	airlin
(m>1)	IC	→	gyroscopic	→	gyroscop
(m>1)	ABLE	→	adjustable	→	adjust
(m>1)	IBLE	→	defensible	→	defens
(m>1)	ANT	→	irritant	→	irrit
(m>1)	EMENT	→	replacement	→	replac
(m>1)	MENT	→	adjustment	→	adjust
(m>1)	ENT	→	dependent	→	depent
(m>1)	and (*S or *T)) ION	→	adoption	→	adopt
(m>1)	OU	→	homologou	→	homolog
(m>1)	ISM	→	communism	→	commun
(m>1)	ATE	→	activate	→	activ
(m>1)	ITI	→	angulariti	→	angular
(m>1)	OUS	→	homologous	→	homolog
(m>1)	IVE	→	effective	→	effect
(m>1)	IZE	→	bowdlerize	→	bowdler

The suffixes are now removed. All that remains is a little tidying up.

Step 5 a

(m>1) E	→	probate	→	probat
		rate	→	rate
(m=1 and not *o) E	→	cease	→	ceas

Step 5 b

(m>1 and *d and *L)	→	single letter		control
	→	controll	→	
		roll	→	roll

The algorithm is careful not to remove a suffix when the stem is too short, the length of the stem being given by its measure, m. There is no linguistic basis for this approach. It was merely observed that m could be used quite effectively to help decide whether or not it was wise to take off a suffix. For example, in the following two lists:

list A	list B
RELATE	DERIVATE
PROBATE	ACTIVATE
CONFLATE	DEMONSTRATE
PIRATE	NECESSITATE
PRELATE	RENOVATE

-ATE is removed from the list B words, but not from the list A words. This means that the following pairs DERIVATE/DERIVE, ACTIVATE/ACTIVE, DEMONSTRATE/DEMONSTRABLE, NECESSITATE/NECESSITOUS, will conflate together. The fact that no attempt is made to identify prefixes can make the results look rather inconsistent. Thus PRELATE does not lose the -ATE, but ARCHPRELATE becomes ARCHPREL. In practice this does not matter too much, because the presence of the prefix decreases the probability of an erroneous conflation.

Complex suffixes are removed bit by bit in the different steps. Thus GENERALIZATIONS is stripped to GENERALIZATION (Step 1), then to GENERALIZE (Step 2), then to GENERAL (Step 3), and then to GENER (Step 4). OSCILLATORS is stripped to OSCILLATOR (Step I), then to OSCILLATE (Step 2), then to OSCILL (Step 4), and then to OSCIL (Step 5). In a vocabulary of 10,000 words, the reduction in size of the stem was distributed among the steps as follows:

Suffix stripping of a vocabulary of	10,000 words
Number of words removed in step 1:	3597
Number of words removed in step 2:	766
Number of words removed in step 3:	327
Number of words removed in step 4:	2424
Number of words removed in step 5:	1373
Number of words not reduced:	3650

The resulting vocabulary of stems contained 6370 distinct entries. Thus the suffix stripping process reduced the size of the vocabulary by about one third (Porter, 1997).

References

Adjei, O. and Mrozek, Z. (1999) Robot Vision: Real time identification of simulated machine parts using features modelled from the Radon space. *Proceedings of the fourth International Conference on System Simulation and Scientific Computing*, BICSC, (pp. 279-284). Beijing, China.

Adjei, O. and Vella, A. (2000) Recognition of human faces based on fast computation of circular harmonic components. *Proceedings of the third International Conference on Multimodal Interfaces*, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, (pp. 160-167). Beijing, China.

Baeza-Yates, R. and Ribeiro-Neto B. (1999) *Modern information retrieval.* ACM Press, Boston, Massachusetts, Addison Wesley.

Baranyi, A. and Koczy, P. (2001) Saving calculation in information retrieval. *Proceedings of the fourth International Conference on flexible query answering systems* (pp. 337-349). Warsaw, Poland.

Beckwith, R. and Miller G.A. (1990) Implementing a lexical network. *International Journal of Lexicography* 3(4), 302 - 312.

Bessis, N. (2003a) Towards a Homogeneous Status of Communicated Research. *Proceedings of the sixth International Conference on Electronic Theses and Dissertations*, ETD, Berlin, Germany.

Bessis, N. (2003b) Proposing an Automated Method for Refereeing Research Work. *Proceedings of the ninth Annual Conference of Chinese Automation and Computing Society in the UK*, CACSCUK, Luton, UK.

Bezdek, J.C. (1980) A convergence theorem for the fuzzy ISODATA clustering algorithms *IEEE Transactions on pattern analysis and machine intelligence* 2(1), 1-8.

Bezdek, J.C., Hathaway, R.J., Sabin, M.J. and Tucker, W.T. (1987) Conversions theory for fuzzy c-means: counter examples and repairs. *IEEE Transactions on systems, man and cybernetics*, 17, 873-877.

Bookstein, A. (1980) Fuzzy requests: An approach to weighted Boolean searches. *Journal of the American Society for Information Science*, 31(4), 240-247.

Bookstein, A. (1986) Probability and fuzzy set applications to information retrieval. *Annual review of information science and technology*, 29, 117-151.

Bordogna, G. and Pasi, G. (1993) A fuzzy linguistic approach generalizing Boolean information retrieval. *Journal of the American Society for Information Science*, 44(2), 70-82.

Bordogna, G. and Pasi, G. (2001) Modelling vagueness in information retrieval. *Proceedings of the fourth International Conference on flexible query answering systems* (pp. 207-241). Warsaw, Poland.

Brandow, R., Mitze, K. and Rau, L. (1995) Automatic condensation of electronic publications by sentence selection. *Information Processing and Management,* 31, 675-685.

Bratley, P., Fox, B.L. and Schrage, L. E. (1983) *A Guide to Simulation.* Springer-Verlag, New York, New York.

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1-7), 107-117.

Brooks, F.P. and Iverson, K.E. (1963) *Automatic Data Processing.* New York, John Wiley and Sons.

Brule, J. (1986) *Artificial intelligence.* Blue Ridge Summit, Pennsylvania, Tab Books.

Buell, D.A. (1981) A general model of query processing in information retrieval. *Information processing and management,* 17(5), 249-262.

Busetta, P., Serani, L., Singh, D. and Zini, F. (2001) Extending multi-agent cooperation by overhearing. *Proceedings of the sixth International Conference on Cooperative Information Systems,* (pp. 40-52). Trento, Italy.

Chakrabarti, S., Dom, B., Agrawal, R. and Raghavan P. (1998) Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal* 7(3), 163-178.

Chen, J., Mikulcic, A. and Kraft, D.H. (2000) An integrated approach to information retrieval with fuzzy clustering and fuzzy inferencing. In Pons, O., Vila, M.A. and Kacprzyk, J., editors, *Knowledge management in fuzzy databases*, Heidelberg, Germany, Physica Verlag.

Cignoli, R.L.O., D'Ottaviano, I.M.L. and Mundici, D. (2000) *Algebraic foundations of Many-valued Reasoning*, Dordrecht, Netherlands, Kluwer Academic Publishers.

Cole, J.I., (Eds.) (2003) *The UCLA Internet Report Surveying the Digital Future*, Los Angeles: University of California Los Angeles Centre for Communication Policy.

Cooper, W.S. (1988) Getting beyond Boolean. *Information processing and management*, 24, 243-248.

Cleverdon, C.W. (1972) On the inverse relationship of recall and precision. *Journal of Documentation* 23, 195-201.

Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A. (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), 391-407.

Devlin, K. (1998) *Goodbye Descartes: The end of logic and the search for a new cosmology of the mind*. New York, John Wiley and Sons.

Donnelly, P.G. (1998) *A domain based approach to natural language modelling,* PhD thesis, Queen's University of Belfast, Belfast, UK.

Eastman, C.M. and Jansen, B.J. (2003) Coverage, relevance, and ranking: The impact of query operators on web search engine results. *ACM Transactions on Information Systems* 21(4), 383- 411.

Edmundson, H.P. (1969) New methods in automatic abstracting. *Journal of the Association for Computing Machinery* 16(2), 264-228.

Estall, C. and Smith, F. J. (1984) Shared processing with an advanced intelligent terminal. *Proceedings of the seventh annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 153-166). Cambridge, UK.

Fairthorne, R.A. (1961) *Towards information retrieval.* London, Butterworths.

Forsyth, R.S. (1995) *Stylistic Structures,* PhD thesis, University of Nottingham, Nottingham, UK.

Frakes, W.B. and Baeza-Yates, R. (1992) *Information retrieval data structures and algorithms,* Englewood Cliffs, New Jersey, Prentice Hall.

Freeman, W.J. (2000) A neurobiological interpretation of semiotics: meaning, representation and information. *Information Sciences,* 124(1-4), 93-102.

Fuhr, N. and Buckley, C. (1991) A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems (TOIS)* 9(3), 223-248.

Fum, D., Guida, G. and Tasso, C. (1985) Evaluating importance: A step towards text summarization. *Proceedings of the International Joint Conference on Artificial Intelligence*, (pp. 840-844). Los Altos, California, USA.

Ganesan, P., Garcia-Molina, H. and Widom, J. (2003) Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems (TOIS)* 21(1), 64-93.

Gedeon, T.D. and Mital V. (1991) Information retrieval using a neural network integrated with hypertext. *Proceedings of the fourth International Conference on neural networks*, Singapore.

Gelsinger, P., Gargini, P., Parker, G. and Yu, A. (1989) Microprocessors circa 2000. *IEEE Spectrum*, 26(10), 43-47.

Hajek, P. (1998) *Metamathematics of Fuzzy Logic*. Dordrecht, Netherlands, Kluwer Academic Publishers.

Harabagiu, S.M. and Moldovan D.I. (1997) TextNet a text-based intelligent system. *Journal for Natural Language Engineering* 3, 171-190.

Hartley, R.J., Keen, E.M., Large, J.A. and Tedd, L.A. (1990) *Online searching: Principles and practice*. London, Bowker Saur.

Hopgood, A. (1993) *Knowledge based systems for engineers and scientists*. Boca Raton, Florida, CRC Press.

Hopgood, A. (2001) *Intelligent systems for engineers and scientists*. Boca Raton, Florida, CRC Press.

Hudson, R. (1995) *Word Meaning, London*, UK, Routledge.

Jacobs, P.S. & Rau, L.F. (1990) SCISOR: Extracting information from on line news. *Communications of the ACM* 33(11), 88-97.

Joachims, T. (1998) Text categorization with support vector machines: learning with many relevant features. *Proceedings of European Conference on Machine Learning*, (pp. 137-142). Chemnitz, Germany.

Johnson, F.C., Paice, C.D., Black, W.J. & Neal, A.P. (1993) The application of linguistic processing to automatic abstract generation. *Journal of Documentation and Text Management* 1(3), 215-241.

Kamel, M., Hadfield, B. and Ismail, M. (1990) Fuzzy query processing using clustering techniques. *Information processing and management* 26, 279-293.

Kehoe, C. and Pitkow, J. (1996) Surveying the territory. *World Wide Web Journal,* 1(3), 77-84.

Kobayashi, I., Chang, M.S. and Sugeno, M. (2002) A study on meaning processing of dialogue with an example of development of travel consultation system. *Information Sciences*, 144(1-4), 45-74.

Knuth, D.E. (1997) *The art of computer programming. Volume 1 Fundamental algorithms, third edition*, Boston, Massachusetts, Addison Wesley.

Kosko, B. (1999) *The fuzzy future*. New York, Harmony books.

Kowalski, G. (1997) *Information retrieval systems: Theory and implementation.* Dordrecht, Netherlands, Kluwer academic publishers.

Kraft, D.H. and Chen, J. (2001) Integrating and extending fuzzy clustering and inferencing to improve text retrieval performance. *Proceedings of the fourth International Conference on flexible query answering systems,* Warsaw, Poland.

Lancaster, F.W. (1978) *Toward Paperless Information Systems*. New York, Academic Press.

Lawrence, S., Giles, C. & Bollacker K. (1999) Digital libraries and autonomous citation indexing. *IEEE Computer* 32(6), 67-71.

Lawrence, S. & Giles, C. (1999) Accessibility of information on the web. *Nature* 400(6740), 107-109.

Lawrence, S. & Giles, C. (1998) Searching the world wide web. *Science* 280(5360), 98-100.

L'Ecuyer, P. (1988) Efficient and portable combined random number generators. *Communications of the ACM* 31(6), 742-751.

L'Ecuyer, P. (1990) Random numbers for simulation. *Communications of the ACM* 33(10), 85-97.

Lehnert, W.G. (1983) Narrative complexity based on summarization algorithms. *Proceedings of the International Joint Conference of Artificial Intelligence*, (pp. 713-716). Karlsruhe, Germany.

Lewis, D.D. (1992) An evaluation of phrasal and clustered representations of a text categorisation task. *Proceedings of the ACM SIGIR Conference on research and development in information retrieval*, (pp. 37-50). New York, New York, ACM Press.

Li, Y.H. and Jain, A.K. (1998) Classification of text documents. *The Computer Journal* 41(8), 537-546.

Liddy, E., Paik, W. and Yu, E.S. (1994) Text categorization for multiple users based on semantic features from a machine readable dictionary. *ACM Transactions on Information Systems (TOIS)* 12(3), 278-295.

Luhn, H.P. (1958) The automatic creation of literature abstracts. *IBM Journal of Research Development* 2, 159-165.

Lukasiewicz, J. (1951) *Aristotle's syllogistic from the standpoint of modern formal logic.* Oxford, Clarendon Press.

Lukasiewicz, J. (1966) *Elements of mathematical logic.* Oxford, Pergamon Press.

Lyman, P. and Varian, H.R. (2003) *How much information.* School of Information Management and Systems, University of California, Berkeley.

Martin, J. (1967) *Design of real-time computer systems.* Series in Automatic Computation. Englewood Cliffs, New Jersey, Prentice Hall.

Mani, I. and Bloedorn, E. (1997) Multi-document summarization by graph search and matching. *Proceedings of the AAAI National conference on Artificial Intelligence.* Cambridge, Massachusetts, USA.

Mauldin, M.L. (1991) *Conceptual information retrieval – A case study in adaptive partial parsing*, Boston, Massachusetts: Kluwer Academic Publishers.

McMahon, J. and Smith, F.J. (1998) A Review of Statistical Language Processing Techniques. *Artificial Intelligence Review*, 12(5), 347 – 391.

Mihalcea, R. & Moldovan D.I. (2001) Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation. *International Journal on Artificial Intelligence Tools* 10(1-2), 5-21.

Mooers, C.N. (1952) Information retrieval viewed as temporal signalling. *Proceedings of the International Conference of Mathematicians*, American Mathematical Society (pp. 572-573). Cambridge, Massachusetts, USA.

Murray-Rust, P. and Rzepa, H. S. (1999) Chemical markup Language and XML Part I. Basic principles. *Journal of Chemical Information and Computer Sciences*, 39(6), 928 - 942.

Murray-Rust, P. and Rzepa, H.S. (2002a) Scientific publications in XML - towards a global knowledge base. *Data Science Journal* 1(1), 84-98.

Murray-Rust, P. and Rzepa, H. S. (2002b) Markup Languages- How to structure chemistry related documents. Chemistry Intl. 24(4), 9-13.

Negnevitsky, M. (2001) *Artificial Intelligence: A Guide to Intelligent Systems*. Boston, Massachusetts, Addison Wesley.

Ogawa, Y., Morita, T. and Kobayashi, K. (1991) A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy sets and systems*, 39(2), 163-179.

Paice, C. (1990) Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management* 26(1), 171-186.

Paice, C. & Jones, P. (1993) The identification of important concepts in highly structured technical papers. *Proceedings of the ACM SIGIR Conference on research and development in information retrieval*, (pp. 69-78). New York, New York, USA.

Petratos, P. (2003) A polythematic real-time synergistic hybrid data telecommunication system for scientific research with bidirectional fuzzy feedback peer review by expert referees. *Data Science Journal* 2(4), 47-58.

Petratos, P. and Chen, L. (2002) A note on bidirectional fuzzy logic. *Proceedings of the North American Fuzzy Information Processing Society Conference*, New Orleans, Louisiana, USA.

Petratos, P., Chen, L., Wang, P. and Forsyth, R. (2002) A Bi-directional Fuzzy Logic Theory: The Generalized Knuth's Triadic Logic for Information Retrieval. *Proceedings of the IEEE Systems Man and Cybernetics Conference*, Hammamet, Tunisia.

Porter, M.F., (1997) An algorithm for suffix stripping. In Sparck Jones, K. and Willett, P., Editors, *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Rasmussen, E. (1992) Clustering algorithms. In Frakes, W.B. and Baeza-Yates, R., Editors, *Information retrieval data structures and algorithms*. Englewood Cliffs, New Jersey, Prentice Hall.

Rau, L.F. and Jacobs, P.S. (1991) Creating segmented databases from free text for text retrieval. *Proceedings of the ACM SIGIR Conference on research and development in information retrieval*, (pp. 337-346). New York, New York, USA.

Rau, L.F., Jacobs, P.S. and Zernik, U. (1989) Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing and Management* 25(4), 419-428.

Riloff, E. (1995) A corpus based approach to domain specific text summarization. In Endres-Niggemeyer, B., Hobbs, J. and Sparck Jones, K., (Eds), *Proceedings of Conference on summarizing text for intelligent communication*, (pp. 69-84). Dagstuhl, Germany.

Riloff, E. & Lehnert, W.G. (1994) Information extraction as a basis for high precision text classification. *ACM Transactions on Information Systems (TOIS)* 12(3), 296-333.

Rush, J.E., Salvador, R. & Zamora, A. (1971) Automatic abstracting and indexing. *Journal of the American Society for Information Science* 22(4), 260-274.

Russell, S.J. and Norvig, P. (2002) *Artificial Intelligence: A Modern Approach (2nd Edition)*. Englewood Cliffs, New Jersey, Prentice Hall.

Rogers, D. (1995) *The Bodleian Library and its Treasures 1320-1700*. Oxon, Aidan Ellis.

Salton, G. (1989) *Automatic text processing*. Boston, Massachusetts, Addison Wesley.

Salton, G., Allan, J., Buckley, C. and Singhal, A. (1997) Automatic analysis, theme generation and summarization of machine readable texts. In Sparck Jones, K. and Willett, P., Editors, *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Salton, G. and Buckley, C. (1997) Improving retrieval performance by relevance feedback. In Sparck Jones, K. and Willett, P., Editors, *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Salton, G. and Buckley, C. (1988) Term weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5), 513-523.

Salton, G. and Lesk, M.E. (1997) Computer evaluation of indexing and text procesing. In Sparck Jones, K. and Willett, P., Editors, *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Salton, G. and McGill, M.J. (1997) The SMART and SIRE experimental retrieval systems. In Sparck Jones, K. and Willett, P., Editors, *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Salton, G., Wong, A. and Yang, C.S. (1997) A vector space model for automatic indexing. In Sparck Jones, K. and Willett, P., Editors, *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Schapire, R.E. and Singer, Y. (2000) BoosTexter: a boosting based system for text categorisation. *Machine Learning* 39 2/3, 135-168.

Smith, F.J. and Linggard, R.J. (1982) Information retrieval by voice input and output. *Proceedings of the fifth ACM conference on research and development in information retrieval*, (pp. 275-288). Berlin, Germany.

Smith, F. J. and Clotworthy, C. J. (1988) A Statistical Study in Word Recognition. *Proceedings of the 4th International Conference on Pattern Recognition*, Springer-Verlag, (pp. 203-215). London, UK.

Smith, F. J. and Devine K. (1985) Storing and retrieving word phrases. *Information Processing and Management: an International Journal*, Volume 21 Issue 3, 215-224.

Sparck Jones, K. (1995) Discourse modelling for automatic summaries. In Hajicova, E., Cervenka, M., Leska, O. and Sgali, P. (Editors) *Prague Linguistic Circle Papers* 1, 201-227.

Sparck Jones, K. (1997) Search term relevance weighting given little relevance information. In Sparck Jones, K. and Willett, P., Editors, *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Sparck Jones, K., Jones, G.J.F., Foote, J.T. and Young, S.J. (1997) Experiments in spoken document retrieval. In Sparck Jones, K. and Willett, P., Editors, *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Sparck Jones, K. and Willett, P. (1997) *Readings in information retrieval*, San Francisco, Morgan Kaufmann Publishers.

Van Rijsbergen, C.J. (1979) *Information Retrieval 2nd edition*, London, UK, Butterworths.

Van Rijsbergen, C.J. (1986) A non classical logic for information retrieval. *Computer Journal*, 29(6), 481-485.

Van Rijsbergen, C.J. and Lalmas, M. (1996) An information calculus for information retrieval. *Journal of the American Society of Information Science*, 47(5), 385-398.

Wagner, C. and Turban, E. (2002) Are intelligent e-commerce agents partners or predators? *Communications of the ACM* 45(5), 84-90.

Wen, J., Nie, J. and Zhang, H. (2002) Query clustering using user logs. *ACM Transactions on Information Systems (TOIS)* 20(1), 59-81.

Wichmann, B.A and Hill, I.D. (1982) An efficient and portable pseudo-random number generator. *Applied Statistics* 31, 188-190.

Witten, I.H., Moffat, A. and Bell, T.C. (1999) *Managing gigabytes, compressing and indexing documents and images*. San Francisco, Morgan Kaufmann Publishers.

Wong, S.K.M. and Yao Y.Y. (1990) Query formulation in linear retrieval models. *Journal of the American Society for Information Science* 41, 334-341.

Xu, J. and Croft, W.B. (1996) Query expansion using local and global document analysis. *Proceedings of the nineteenth ACM-SIGIR Conference on research and development in information retrieval*, (pp. 4-11). Zurich, Switzerland.

Yager, R. and Kacprzyk, J. (1997) *The ordered weighted averaging operators: theory and applications*. Dordrecht, Kluwer academic publishers.

Yang, Y. (1994) Expert network: effective and efficient learning from human decisions in text categorisation and retrieval. *Proceedings of the ACM SIGIR*

Conference on research and development in information retrieval, (pp. 13-22). Dublin, Ireland.

Yarowsky, D. (1992) Word sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of the International Conference on Computational Linguistics*, (pp. 454-460). Nantes, France.

Zadeh, L. (1965) Fuzzy sets. *Information and control* (8), 338-353.

Zobel, J. and Moffat, A. (1998) Exploring the similarity space. *ACM-SIGIR Forum,* 32(1), 18-34.

Zobel, J. (1998) *How reliable are the results of large scale information retrieval experiments?* Technical Report 98-1, Department of Computer Science, RMIT, Melbourne, Australia.