Information Retrieval Systems: A Human Centered Approach

Panagiotis Petratos
California State University, Stanislaus, CA, USA

ppetratos@csustan.edu

Abstract

In this article a human centered approach for Information Retrieval IR systems is offered. The focus of this article is on human users of IR systems, who are given the opportunity to offer enhanced relevance feedback. A novel approach is offered to capture the back-propagated feedback of human users during IR sessions. According to this approach the IR system is developed to accept human bidirectional relevance feedback, which is utilized to further enhance the IR process with the aim of improving IR effectiveness.

Keywords: Information Retrieval Systems, Human Computer Interaction, Bidirectional Relevance Feedback.

Introduction

Top Universities offer a high quality, in depth education to Engineers and Computer Scientists, regarding the intricate technical details of electronics, microprocessors, computing systems design and algorithms development. These graduates subsequently practice design methodologies that are overly focused on every technical specification of the machine. Often, they are so much engrossed in the technical details that the human parameter is often neglected, or, worse, simply forgotten.

It is no secret that a good system design is frequently elusive, despite the amount of resources invested in a project. The examples of well-known, high-tech corporations caught by surprise, when the unanticipated market failure of their products occurs, despite the very good technical specifications of their machines, are numerous.

For instance, the author is currently using an ultra portable notebook computer, designed and manufactured by a well known, high-tech corporation. This particular notebook computer is often rated among the best by independent testing laboratories for mobile application performance and battery life.

However, the manufacturer mysteriously incorporated a left-handed mouse for everyone, regardless of the user's handedness.

Thus, the mouse input connection is positioned on the left side of the computer and the optical DVD drive is positioned on the right side of the computer. It would seem to an unsuspecting observer that the designer intentionally arranged the computer that way. Unfortunately, there is no right-handed version of this computer.

Editor: Eli Cohen

It often seems, despite the fact that these admirable machines are designed for human users, their convenience, ease of use and simple practicality are typically the last thoughts in the minds of the designers. Naturally, computing information systems are no exception.

IR systems, in particular, are the focus of this article. They are equally susceptible to *ease of use* issues, as well as, to modern *overabundance of information* issues, which are discussed next in the motivation section.

Motivation

What is the Problem?

Current conditions

In this section the motivation for this research is explained. The exploration of the current developments and advances of information technology and a glimpse into the immediate future directions of computing research and development reveal the conditions, which influence and affect IR systems.

A study by Lyman and Varian reports that 800 MB of information were produced in 2003 for every living human being on earth (Lyman & Varian, 2003). More recently another study reports that the available electronic information online on the World Wide Web exceeds 11.5 billion pages (Gulli & Signorini, 2005). According to other recent estimates over 90% of information produced is digital (Varian, 2005). These are very significant figures which clearly show the ever-flourishing *overabundance of information* which is continuously invigorated by technology.

These conditions lead the observer to the definitions of the exact problems facing IR system designers for the present and in the future. Naturally, the problems can also be seen, by inquisitive minds, as opportunities for further research. The two primary research areas that require the attention and endeavor of IR researchers are the *ease of use,* often stemming from ineffective human computer interaction, and the modern *overabundance of information* (Petratos, 2006).

New inventions augment the problem

The most prominent and most influential new inventions, which also won the Millennium Technology Prize, are the World Wide Web and the bright blue, green and white light emission diodes, as well as the invention of the blue laser (MPF, 2006).

These new inventions have the potential to significantly increase the available electronic information, in optical storage media, in fiber-optic telecommunications, as well as online.

At this point, it is noteworthy to indicate that all inventions that were awarded the world's greatest technology prize are the most significant contributors to the ever flourishing *overabundance of information*. This state of events is most commonly known as the information overload problem.

The Attempt to Offer a Solution

IR Systems

Modern IR computing systems

The IR field is initially developed in libraries (Staikos, 2000), and the first attempts at locating information of interest by electronic means fledge (Grossman & Frieder, 2004). In the next IR

progress stage, development efforts are focused on probabilistic (Chaudhuri, Das, Hristidis, & Weikum, 2006) and statistical methods.

During the 1990s, the World Wide Web was invented. The document collections contained in the World Wide Web overshadow in size all other collections, whilst they continue to increasingly grow to the present.

Naturally, the information carried through the Internet is in much greater quantities than the World Wide Web document collections. The World Wide Web is merely one of the services offered by the Internet architecture; others include email (Petratos & Gleni, 2006), telnet, file transfer, chat, etc.

As the reader would expect, the problem is not only the public Internet. The popularity, scalability, low cost, and open architecture of the World Wide Web have made it a preferred publishing medium, as well as a digital library of choice for many organizations (Bar-Ilan, 2004).

The Google mini and the GB-8008 are specially developed search appliances for enterprises (Google, 2006). Currently, the GB-8008 is capable of searching up to thirty million documents.

However, if an organization archives collections with size even greater than thirty million documents, this enterprise search appliance can be customized to manage the additional number of documents for an additional fee, per case.

Quite a few search appliances have been sold to numerous well-known international corporations, as well as distinguished Universities. Consequently, all of these organizations also contribute, in an indirect way, to the modern *overabundance of information*.

A simple IR example

As a result, now and for the foreseeable future, all of us shall be submerged in oceans of information. For example, the major search engines, such as Yahoo and Google, index billions of documents. Even a simple query issued to them often returns over one million results.

For instance, consider the simple following IR example (Figure 1). An enquiry, issued to Google about "*multiple sclerosis therapies*" returns more than 2.54 million documents in the answer set. Fortunately, Google is fast, as it takes only 0.20 seconds for this IR system to return the full answer set.

Currently, there are only two widely approved medicines by the USA *Food and Drug Administration FDA* for *multiple sclerosis MS* patients at early stages of the disease. The medicines are Interferon beta variants, i.e. interferon beta-1a, 1-b, etc. given to the patients with periodic injections, monthly or biweekly, and Glatiramer acetate, given to the patients with daily injections.

Mitoxantrone and Natalizumab are two new medicines that have been recently approved by the *FDA* for use in specific progressive cases of patients at an advanced stage of the disease.

Additional, new, oral medicines, instead of injections, have been recently discovered, such as, Fingolimod and Laquinimod, and still are in Phase I or II of studies. This of course, means that they are still years away from final *FDA* wide approval.

In order for someone to find out this simple piece of information about *MS* patients, the searcher must *fully read all* the top result-documents returned in the answer set, and if she does not find this information, she is forced to fully read even more documents, from the subsequent result-lists of the answer set.

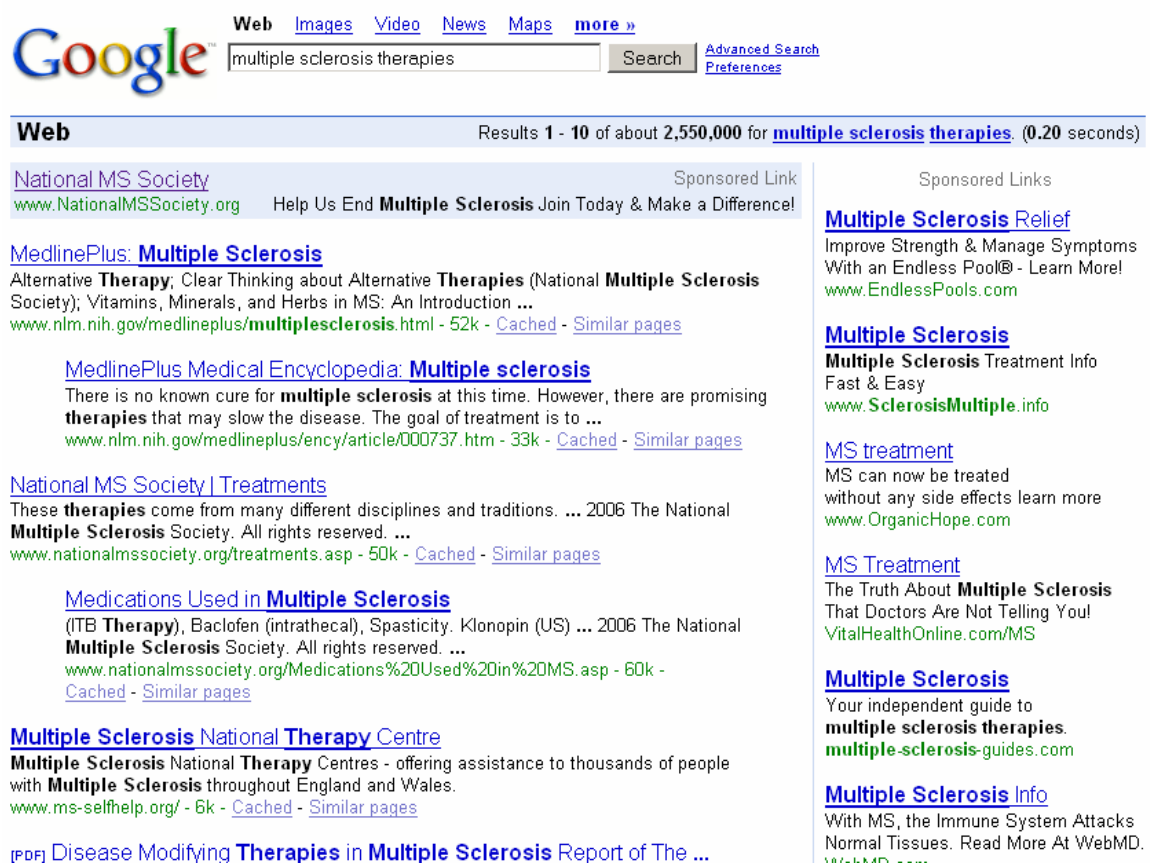


Figure 1. A simple IR example of a query, issued to Google, about "*multiple sclerosis therapies*".

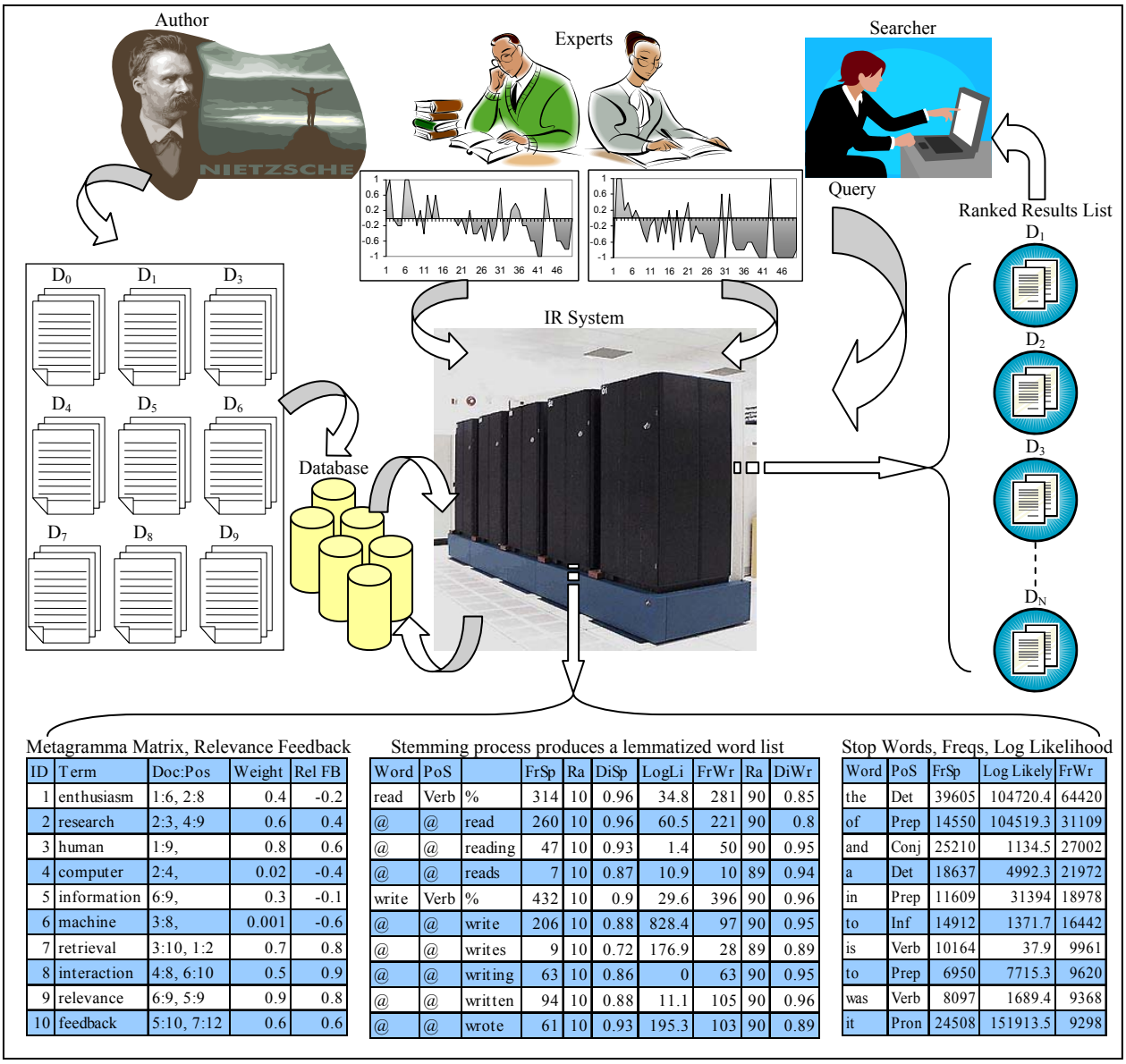
In some other query cases, experiments have shown, there may be information of interest, in a lower ranked result-document, far below the first ten result documents, i.e. 20s, 30s, 40s, etc. returned in the answer set.

As a result, the searcher is often required to fully read a long series of documents, before she is able to extract the desired information. An attempt to offer an alternative strategy, in order to address a few of the aforementioned issues, is the current work.

Functionality of the Experimental IR System

Design

Anacalypse, in Greek means *discovery*. An experimental IR system, called *Anacalypse,* is developed, which attempts to locate information of interest, not only by relying on statistical text processing techniques and incisive text characteristics, but also by accepting human bidirectional relevance feedback, in order to improve IR effectiveness.



Metagramma Matrix, Relevance Feedback

ID	Term	Doc:Pos	Weight	Rel FB
1	enthusiasm	1:6, 2:8	0.4	-0.2
2	research	2:3, 4:9	0.6	0.4
3	human	1:9,	0.8	0.6
4	computer	2:4,	0.02	-0.4
5	information	6:9,	0.3	-0.1
6	machine	3:8,	0.001	-0.6
7	retrieval	3:10, 1:2	0.7	0.8
8	interaction	4:8, 6:10	0.5	0.9
9	relevance	6:9, 5:9	0.9	0.8
10	feedback	5:10, 7:12	0.6	0.6

Stemming process produces a lemmatized word list

Word	PoS		FrSp	Ra	DiSp	LogLi	FrWr	Ra	DiWr
read	Verb	%	314	10	0.96	34.8	281	90	0.85
@	@	read	260	10	0.96	60.5	221	90	0.8
@	@	reading	47	10	0.93	1.4	50	90	0.95
@	@	reads	7	10	0.87	10.9	10	89	0.94
write	Verb	%	432	10	0.9	29.6	396	90	0.96
@	@	write	206	10	0.88	828.4	97	90	0.95
@	@	writes	9	10	0.72	176.9	28	89	0.89
@	@	writing	63	10	0.86	0	63	90	0.95
@	@	written	94	10	0.88	11.1	105	90	0.96
@	@	wrote	61	10	0.93	195.3	103	90	0.89

Stop Words, Freqs, Log Likelihood

Word	PoS	FrSp	Log Likely	FrWr
the	Det	39605	104720.4	64420
of	Prep	14550	104519.3	31109
and	Conj	25210	1134.5	27002
a	Det	18637	4992.3	21972
in	Prep	11609	31394	18978
to	Inf	14912	1371.7	16442
is	Verb	10164	37.9	9961
to	Prep	6950	7715.3	9620
was	Verb	8097	1689.4	9368
it	Pron	24508	151913.5	9298

Figure 2. A simplified diagram of the architecture of the experimental IR system and some of its internal data structures.

In Figure 2, a simplified diagram of the architecture of the experimental IR system is presented. The database contains a range of collections of documents by various authors. The experimental IR system retrieves documents from the database, according to the queries issued by the searcher.

Subsystems

Also, the experimental IR system is designed with two principle, collaborative, subsystems, which communicate and cooperate during the IR process. One of the subsystems is written in the Python programming language (please see Figure 3).

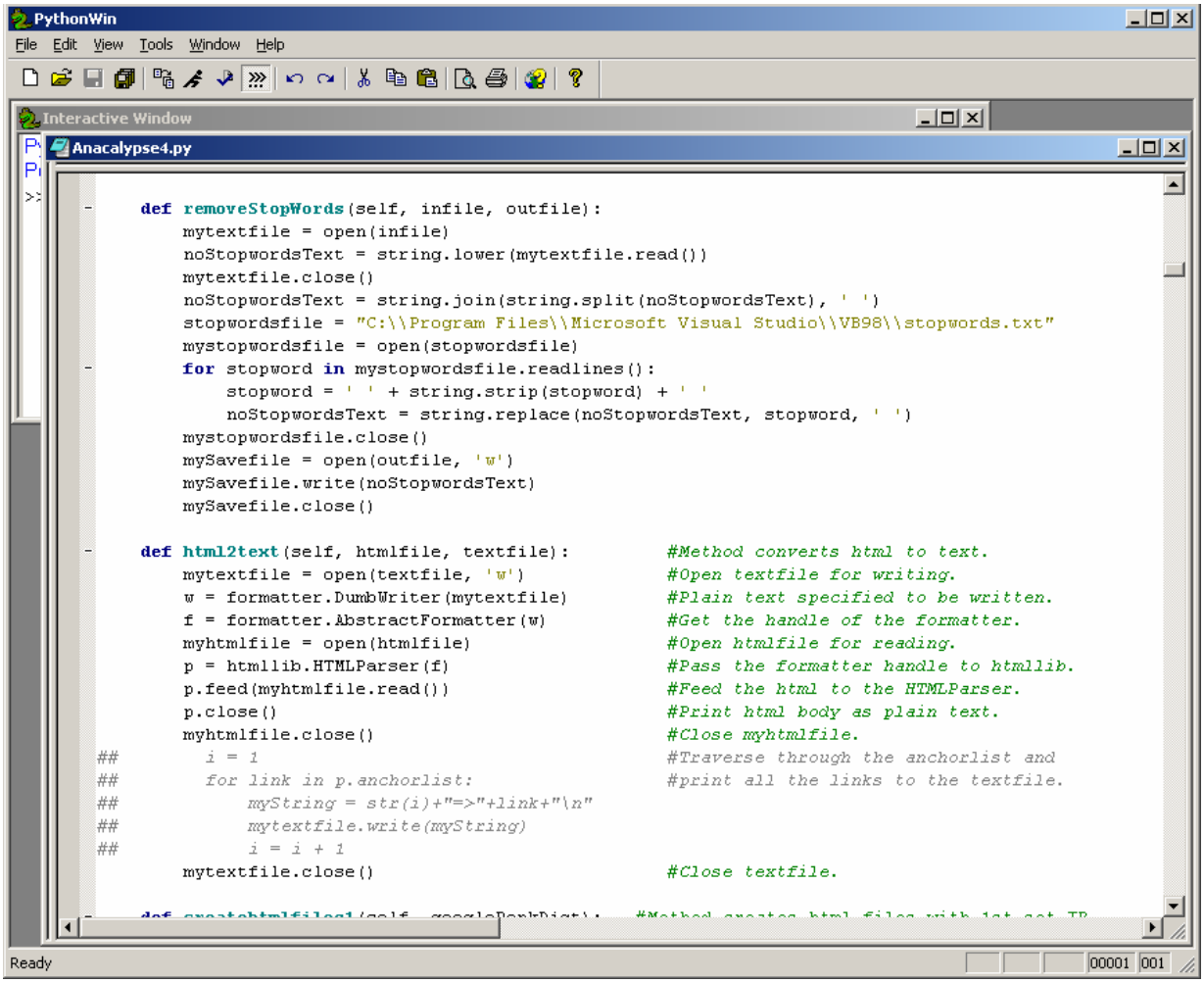


Figure 3. Sample Python code of the experimental IR system for removal of all tags and all stop words.

This inconspicuous subsystem is not visible to the end user, as it performs a lot of the work behind the scenes and is usually referred to as the *back end* subsystem. This Python back end subsystem is responsible for all the text processing tasks, as well as all the statistical computations.

The other subsystem is written in the programming language Visual Basic (please see Figure 4). This conspicuous subsystem is visible to the end user, as it performs a lot of the work in the foreground and is usually referred to as the *front end* subsystem. This front end subsystem is responsible for all the graphical user interface operations, as well as accepting all human input.

Object Functions

The experimental IR system contains various objects and methods capable of a wide range of functions. For instance, there are methods, which can remove all the descriptive tags and extract the pure text of the documents (please see Figure 5). There are methods, which can remove the most common words, from the processed texts (please see Figures 6, 7).

Figure 4. The user interface of the experimental IR system with a sample downloaded document of a Nobel Medicine query.

These frequently common words are often considered not important by scholars, as they carry no serious semantic meaning. However, this work finds some evidence contrary to this belief. Common words can carry an important semantic meaning in a phrase and as shown herein may improve IR effectiveness, when included in the IR process (Petratos, 2004). Semantic term matching is another recent approach to IR (Fang & Zhai, 2006).

Moreover, there are methods, which can perform lemmatization, or stemming (Porter, 1997), extracting the lemmas, or the lexical roots, of the words (please see Figure 7). In addition, there are methods, which can display only the text of documents by hiding their images, figures, or other multimedia contents.

Unsurprisingly, this feature also had an interesting impact on human computer users. Modern human computer users are so accustomed to a graphical, multimedia, informing experience, that they simply denied continuing their IR sessions, without the multimedia contents present.

<table>
<tr>
<td>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html lang="en" xml:lang="en"
xmlns="http://www.w3.org/1999/xhtml">
<head>
<title>Nobelprize.org</title>
<meta http-equiv="Content-Type" content=
"text/html; charset=iso-8859-1" />
<meta name="description" content="Nobelprize.org,
Official web site of the Nobel Foundation" />
<meta name="keywords" content=
"Nobel, Nobelprize, Nobelpriset, Foundation, Prize, Alfred, Museum,
Literature, Physics, Chemistry, Peace, Medicine, Physiology, Economics,
Laureate, Laureates, Winner, Winners, Award, Awards, Science" />
<link href="/css/markup.css" rel="stylesheet" type="text/css" />
<link href="/css/print.css" rel="stylesheet" type="text/css" media="print" />
<!--eri-no-index-->
<script language="javascript" type="text/javascript"
src="/prog/js/sitescripts.js"></script>
<script language="javascript" type="text/javascript"
src="/prog/js/detection_flash.js"></script>
<!--/eri-no-index-->
</head>
<body id="Front">
<!-- Start Header Area -->
<!--eri-no-index-->

<script language="javascript" type="text/javascript"
src="/images/detection/stats_tracker.js"></script>
<noscript><img src="/images/detection/stats_tracker.js?nojs=y"
height="0" width="0" border="0" style="display: none" alt=""
/></noscript>
<!-- Start Header Top -->
<form action="http://search.nobelprize.org/search/nobel/"
method="get" name="search_form" id="search_form">
<ul id="layout_header_top">
<li class="layout_organisation_links"
id="Nobel_Foundation">Nobel Foundation
<li class="layout_organisation_links" id="Nobel_Media"><a
href="/nobelmedia/index.html">Nobel Media
<li class="layout_organisation_links" id="Nobel_Museum"><a
href="/nobelmuseum/index.html">Nobel Museum
<li class="layout_organisation_links" id="Nobel_Peace"><a
</td>
<td>
Nobel Foundation[1]
Nobel Media[2]
Nobel Museum[3]
Nobel Peace Center[4]
Nobel Web[5]

SEARCH[6]
CONTACT US[7]
HOME[8]

nobelprize.org Logo[9]

NOBEL PRIZES[10]
ALFRED NOBEL[11]
PRIZE AWARDERS[12]
NOMINATION[13]
PRIZE ANNOUNCEMENTS[14]
AWARD CEREMONIES[15]
EDUCATIONAL GAMES[16]

Get to know the 2006 Nobel Laureates!

Americans predominated this year, with Nobel Laureates Andrew Fire, Roger Kornberg, John Mather, Craig Mello, George Smoot, and the Economics Laureate Edmund Phelps taking four of the six awards for the first time since 1976.
Literature Laureate, Turkish writer Orhan Pamuk, and the recipients of the Nobel Peace Prize, Muhammad Yunus and the Grameen Bank that he founded in Bangladesh, will join the others heading to Stockholm and Oslo for the December festivities. Read More[17] Registered trademark of the Nobel Foundation

2006 Nobel Prizes

RNAFor Their Discovery of RNA Interference Watch the press conference announcing the 2006 Nobel Prize in Medicine. Read More [18] Kornberg"Well, it's Wonderful News!"
Roger Kornberg was interviewed immediately after
the announcement of the 2006 Nobel Prize in Chemistry. Read More [19] IstanbulThe City and the World Read a two minute summary of the 2006 Nobel Prize in Literature. Read More [20] Per Carlson"A Prize in the Area of Cosmology" Professor Per Carlson explains the 2006 Nobel Prize in Physics in this video. Read More [21] EconomicsA Deepened Understanding Edward S. Phelps's analyses have had a profound impact on economic theory as well as on macroeconomic policy. Read More [22] telephonePeople Can Break Free from Poverty Listen to a short interview with Nobel Peace Prize Laureate Muhammad Yunus. Read More [23]
</td>
</tr>
</table>

Figure 5. Sample of the downloaded document's raw data including all tags and pure text after removal of all tags.

In addition, the front end subsystem is capable of displaying histograms of word lists, the history of past IR sessions, visual representation of web sites on tree-like dendrogram graphs, light-weight web browsers, maximum visibility of documents by dynamically resizing windows after the user shows or hides the view of various graphical user interface components, etc.

English Lexicon

The graphical user interface for the English lexicon assists the user to select the appropriate set of most common words (please see Figure 6). The English lexicon displays an example graph of the few, top, most common words (Gaines, 1989). Although the current work is focused on the English language, the software is designed with a general purpose in mind, so that the experimental IR system can fairly easily accommodate additional languages.

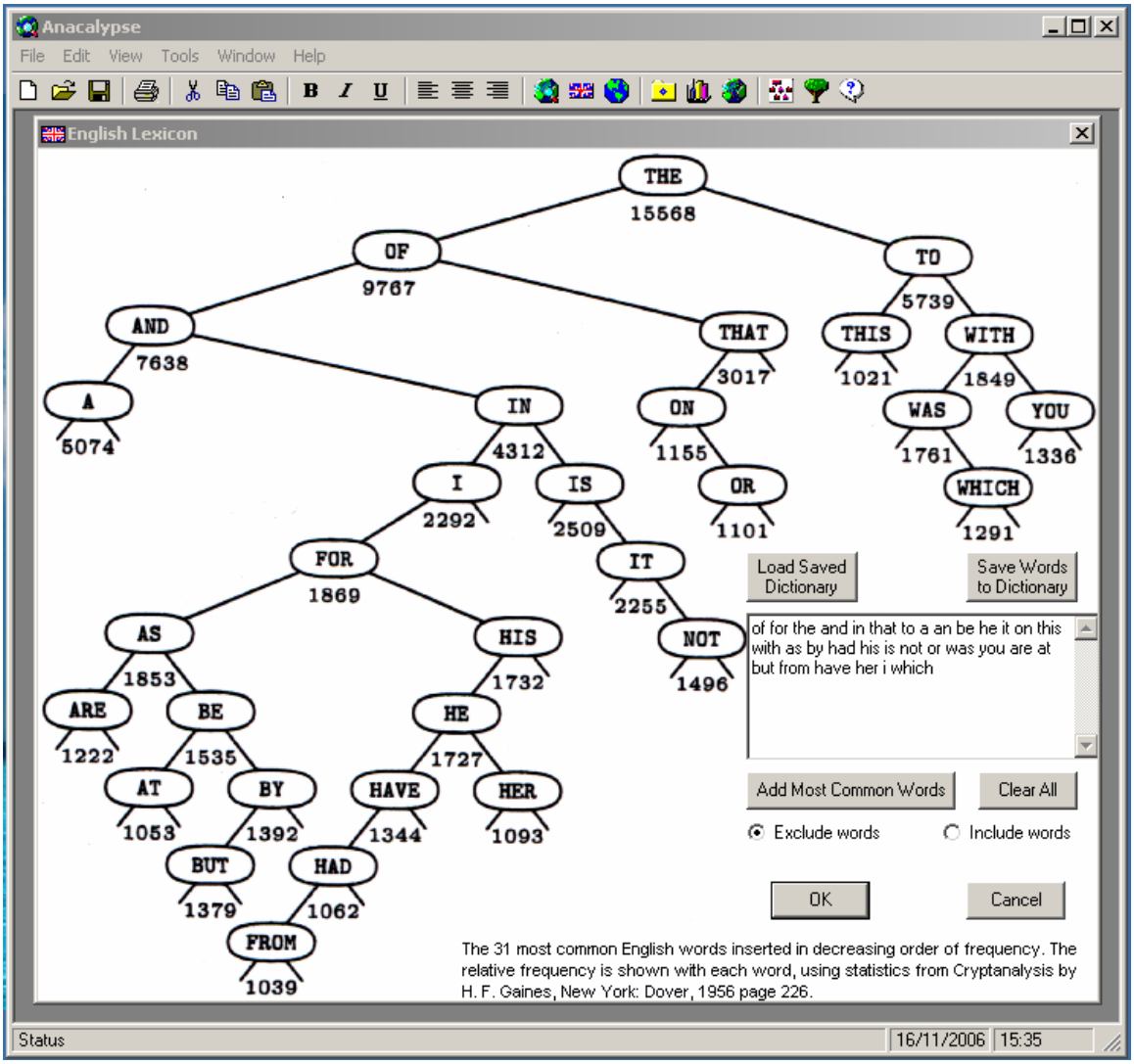


Figure 6. The English lexicon user interface of the experimental IR system with a small sample set of stop words (Gaines, 1989).

Also, the English lexicon user interface provides a mechanism for the user to select, one specific action from a series of choices.

The searcher can select the default set of the most common English words. The searcher can load words from a dictionary previously saved on the hard disk. The user can select to exclude the selected words, or to include the words of her choice. The user can add her personal words to a new dictionary and save it, even in a different language, etc.

Bidirectional Relevance Feedback

Also, a bidirectional fuzzy relevance feedback mechanism is developed, along with the appropriate graphical user interface controls, to give the capability to the IR experimental system to accept relevance feedback, from human users. This mechanism is based on bidirectional fuzzy logic (Petratos, 2003; Petratos, Chen, Wang, & Forsyth, 2002).

Nobel Foundation[1] Nobel Media[2] Nobel Museum[3] Nobel Peace Center[4] Nobel Web[5] SEARCH[6] CONTACT US[7] HOME[8] nobelprize.org Logo[9] NOBEL PRIZES[10] ALFRED NOBEL[11] PRIZE AWARDERS[12] NOMINATION[13] PRIZE ANNOUNCEMENTS[14] AWARD CEREMONIES[15] EDUCATIONAL GAMES[16] 2006 Nobel Laureates! Americans predominated year, Nobel Laureates Andrew Fire, Roger Kornberg, John Mather, Craig Mello, George Smoot, Economics Laureate Edmund Phelps taking awards time 1976. Literature Laureate, Turkish writer Orhan Pamuk, recipients Nobel Peace Prize, Muhammad Yunus Grameen Bank founded Bangladesh, join heading Stockholm Oslo December festivities. Read More[17] Registered trademark Nobel Foundation 2006 Nobel Prizes RNAFor Discovery RNA Interference Watch press conference announcing 2006 Nobel Prize Medicine. Read [18] Kornberg"Well, Wonderful News!" Roger Kornberg was interviewed immediately announcement 2006 Nobel Prize Chemistry. Read [19] IstanbulThe City World Read minute summary 2006 Nobel Prize Literature. Read [20] Carlson"A Prize Area Cosmology" Professor Carlson explains 2006 Nobel Prize Physics video. Read [21] EconomicsA Deepened Understanding Edward S. Phelps's analyses profound impact economic theory macroeconomic policy. Read [22] telephonePeople Break Free Poverty Listen short interview Nobel Peace Prize Laureate Muhammad Yunus. Read [23]	Nobel Foundation[1] Nobel Media[2] Nobel Museum[3] Nobel Peac Center[4] Nobel Web[5] SEARCH[6] CONTACT US[7] HOME[8] nobelprize.org Logo[9] NOBEL PRIZES[10] ALFR NOBEL[11] PRIZE AWARDERS[12] NOMINATION[13] PRIZE ANNOUNCEMENTS[14] AWARD CEREMONIES[15] EDUC GAMES[16] 2006 Nobel Laureates! American predomin year, Nobel Laureat Andrew Fire, Roger Kornberg, John Mather, Craig Mello, Georg Smoot, Econom Laureat Edmund Phelp take award time 1976. Literatur Laureate, Turkish writer Orhan Pamuk, recipi Nobel Peac Prize, Muhammad Yunu Grameen Bank found Bangladesh, join head Stockholm Oslo Decemb festivities. Read More[17] Regist trademark Nobel Foundat 2006 Nobel Prize RNAFor Discoveri RNA Interfer Watch press confer announc 2006 Nobel Prize Medicine. Read [18] Kornberg"Well, Wonder News!" Roger Kornberg wa interview immedi announc 2006 Nobel Prize Chemistry. Read [19] IstanbulTh Citi World Read minut summari 2006 Nobel Prize Literature. Read [20] Carlson"A Prize Area Cosmology" Professor Carlson explain 2006 Nobel Prize Physic video. Read [21] EconomicsA Deepen Understand Edward S. Phelps's analys profound impact econom theori macroeconom policy. Read [22] telephonePeopl Break Free Poverti Listen short interview Nobel Peac Prize Laureat Muhammad Yunus. Read [23]

Figure 7. Sample of the downloaded document's processed text subsequent to removal of all stop words and after applying Porter's stemming algorithm.

The experimental IR system accepts bidirectional relevance feedback, from the human experts, who review the documents of the database and the ranked lists of the results. According to their expert findings, they utilize graphical user interface controls to assign a bidirectional grade to each document in the symmetrical space [-1.0, 1.0].

After the human bidirectional relevance feedback is accepted by the experimental IR system, it is taken into account, in the computation of the weights of the respective vectors, in each document collection matrix. The experimental IR system also computes the various similarity coefficients and compares them.

Representation of Documents

In the experimental IR system, a document can be represented by a multi-dimensional term vector. Each unique term in the document represents a different dimension. Other documents can also be represented by their respective term vectors.

Also, a query can be as well represented by its respective term vector. A further, simple addition is to assign weights to each term-dimension. Multiplicity, or frequency of terms, i.e. six "money" terms, is computed in the term-weight calculation.

Hence, all documents in a particular collection can be represented by their respective term-weight vectors. Another simple normalization, in order for all documents to have the same vector length of terms, is the following.

If a specific document contains an additional term in comparison to another document, then the non-containing document is assigned a zero weight for that term in its term-weight vector.

As a result, the whole collection of documents can be represented by a matrix where one side represents the documents and the other side represents the terms. Naturally, the values of the matrix elements are the assigned weights of each term in the respective document.

This geometrical model, which equally represents documents and queries as vectors in a multi-dimensional space, is called the vector space model (Salton, 1989).

Similarity Functions

This concept is illuminated if the reader considers a diagram in the vector space model. For example, this diagram can be formed by a query vector and two different document vectors. All three vectors have different angles amongst them. If one angle is smaller than the other two, then the Euclidian distance of the specific two vectors forming this angle is the closest.

As a result of the two term vectors, having the closest distance, one from the other, the similarity between the two corresponding documents, that the two term vectors represent, is the greatest. This document similarity can be measured by calculating the cosine of the angle between the two vectors (Salton & McGill, 1997).

In addition to the cosine similarity measure, the experimental IR system is also capable of computing various other similarity measures (please see Table 1 and Figure 8). For instance, the inner product, overlap, Dice, as well as the Jaccard similarity coefficients, can be computed and compared.

Table 1: Similarity functions

NAME	FORMULA
Inner Product	$Sim(D_i, Q_j) = \sum_{x=1}^{m} \left(w_{ix} \cdot w_{jx}\right)$
Cosine	$Sim\left(D_i, Q_j\right) = \frac{\sum_{x=1}^{m} \left(w_{ix} \cdot w_{jx}\right)}{\sqrt{\sum_{x=1}^{m} \left(w_{ix}\right)^2 \cdot \sum_{x=1}^{m} \left(w_{jx}\right)^2}}$
Dice	$Sim\left(D_i, Q_j\right) = \frac{2 \cdot \sum_{x=1}^{m} \left(w_{ix} \cdot w_{jx}\right)}{\sum_{x=1}^{m} \left(w_{ix}\right)^2 + \sum_{x=1}^{m} \left(w_{jx}\right)^2}$
Jaccard	$Sim\left(D_i, Q_j\right) = \frac{\sum_{x=1}^{m} \left(w_{ix} \cdot w_{jx}\right)}{\sum_{x=1}^{m} \left(w_{ix}\right)^2 + \sum_{x=1}^{m} \left(w_{jx}\right)^2 - \sum_{x=1}^{m} \left(w_{ix} \cdot w_{jx}\right)}$

```
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, 0.8709285214433542, 0.82614387318568749, 0.87732985878840808, ..... 0.00000000000000000]
metagram[102] = [0.97785545982754007, 0.9058279975024502, 0.93031688172566296, 0.00000000000000000, ..... 0.00000000000000000]
metagram[103] = [0.95702186853519633, 0.9529990555721367, 0.91599282418456163, 0.88873803717323474, ..... 0.00000000000000000]
metagram[104] = [0.95179208786656222, 0.9119804358867578, 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 8. Sample Metagramma matrix values, similarity coefficients and Spearman correlation statistics

Results

A series of experiments can be carried out with interactive expert relevance feedback in order to categorize information correctly from irrelevant to highly relevant. The queries selected are in the expert's area of expertise in order for the relevance feedback to be accurate. For example, a few of the queries are "Itanium instruction set", "California commercial insurance laws", "Itanium source code porting", "Artificial intelligence algorithms", "Data mining methodologies", etc. A few of the results are displayed in Figures 9, 10.

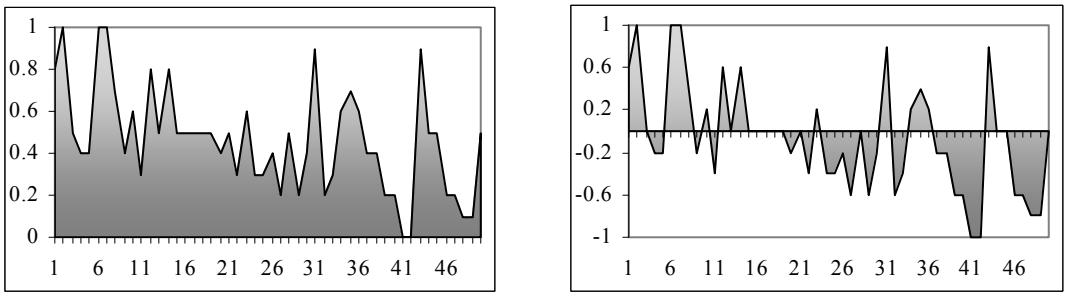


Figure 9. Unidirectional fuzzy supervised training q1 left and Bidirectional fuzzy supervised training q1 right.

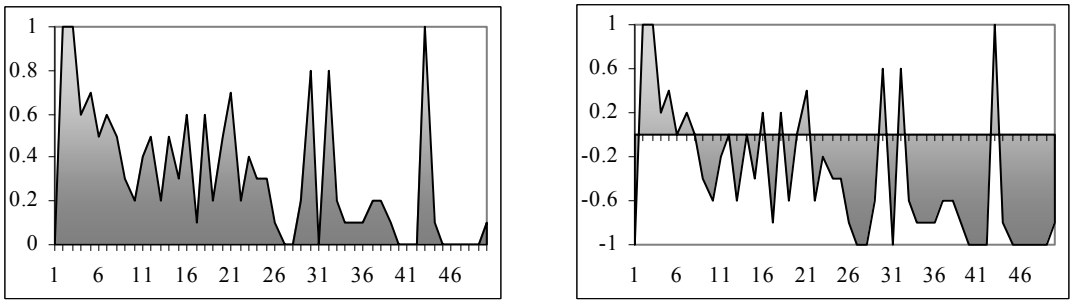


Figure 10. Unidirectional fuzzy supervised training q2 left and Bidirectional fuzzy supervised training q2 right.

The subject matter experts are researchers from the University of Bedfordshire. They are selected in a diverse approach to represent a wide range of demographics (See Table 2).

Table 2: Experts

Expert ID#	Gender	Age	Native Speaker	Expert ID#	Gender	Age	Native Speaker
1	1	0	1	5	0	1	1
2	1	0	0	6	0	0	1
3	1	1	0	7	0	0	0
4	1	1	1	8	0	1	0

The gender value means male=1, female=0, the age value means >=30=1, <30=0, the native speaker value means English=1, other=0. The experiments have two objectives: a) to automatically compute the relevance of unknown documents from a small number of evaluated documents, b) to compare the results of the two IR systems to the expert's relevance standard. The following table and figures summarize the results obtained by processing 2,000 documents (See Table 3).

Table 3: Results summary for 2,000 documents processed.					
Expert ID	**Query**	**Anacalypse correlation**	**Google correlation**	**Difference A-G**	**Anacalypse Change**
1	1	-0.1636	0.3212	-0.4848	-24%
2	2	0.2484	0.4303	-0.1819	-9%
3	3	0.1031	0.6121	-0.509	-25%
4	4	0.6484	-0.1878	0.8362	42%
5	5	0.8545	-0.0424	0.8969	45%
6	6	0.9757	-0.0424	1.0181	51%
7	7	0.5272	-0.0424	0.5696	28%
8	8	0.9878	-0.1393	1.1271	56%
1	9	0.6727	-0.1393	0.812	41%
2	10	0.7091	0.0909	0.6182	31%
3	11	0.7212	0.0909	0.6303	32%
4	12	0.3696	0.4061	-0.0365	-2%
5	13	0.3575	0.6001	-0.2426	-12%
6	14	0.4303	0.6001	-0.1698	-8%
7	15	0.5393	0.6001	-0.0608	-3%
8	16	0.8181	0.9151	-0.097	-5%
1	17	0.4061	0.9151	-0.509	-25%
2	18	0.1515	0.3333	-0.1818	-9%
3	19	-0.0424	-0.0909	0.0485	2%
4	20	-0.0061	-0.0909	0.0848	4%
Average	Values:	0.46542	0.256995	0.208425	10%

The results of the experimental IR system are compared with the expert relevance standard. The results of the commercial IR system are also compared with the expert relevance standard. The results of the two IR systems are compared using the Spearman rank correlation

Figures 11 and 12 show the scatter plots comparing the experimental IR system to the expert relevance standard and contrasting the commercial IR system to the expert relevance standard.

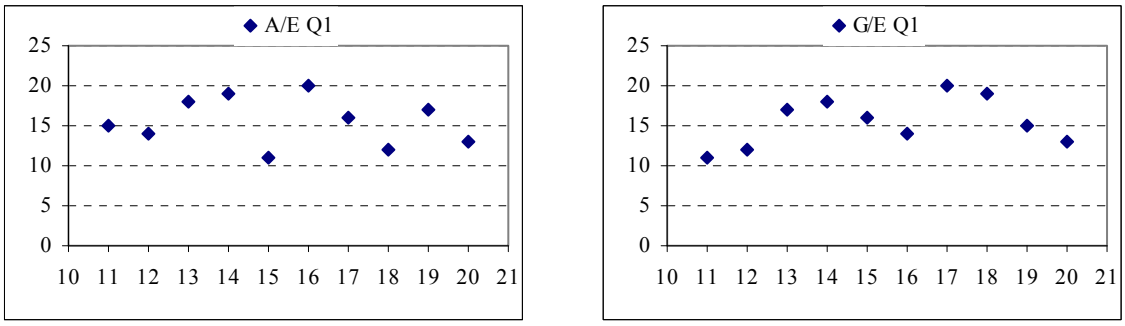


Figure 11. Scatter plots for q1 Anacalypse/Expert (left), Google/Expert (right) correlations.

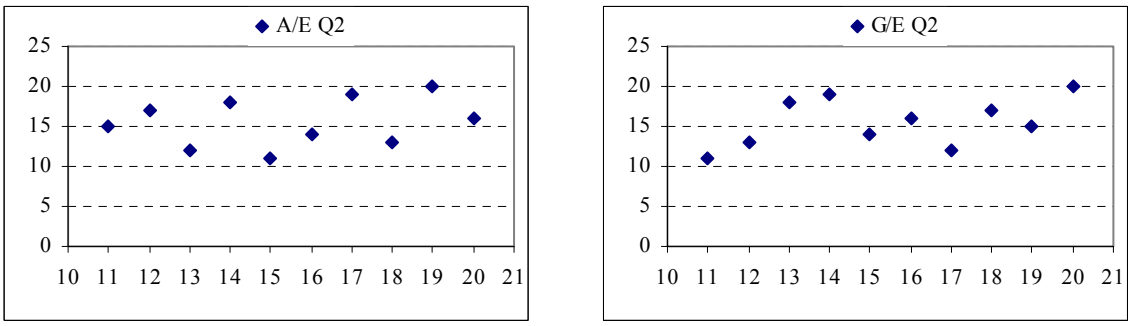


Figure 12. Scatter plots for q2 Anacalypse/Expert (left), Google/Expert (right) correlations.

Conclusion

Finally, the purpose of this work is to offer an alternative IR strategy with some payoff for the searcher. This alternative IR strategy involves and engages the human element in the IR process. The results show some benefits, ten percent average improvement, from using this alternative IR strategy. Other interesting findings are the potential usefulness of the common words, if included in the IR process, and the complete user-reliance to multimedia human computer interfaces.

References

Bar-Ilan, J. (2004). An outsider's view on "topic-oriented blogging". *Proceedings of the 13th international World Wide Web conference on Alternate track papers and posters*, ACM, New York, NY.

Chaudhuri, S., Das, G., Hristidis, V., & Weikum, G. (2006). Probabilistic information retrieval approach for ranking of database query results. *ACM Transactions on Database Systems (TODS), 31*(3), 1134 – 1168.

Fang, H., & Zhai, C. (2006). Semantics: Semantic term matching in axiomatic approaches to information retrieval. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR '06,* ACM, New York, NY.

Gaines, F. H. (1989). *Cryptanalysis.* London, England: Dover Publications.

Grossman, D. & Frieder, O. (2004). *Information retrieval: Algorithms and heuristics.* New York, NY: Springer.

Google. (2006). *Google enterprise solutions*. Retrieved from http://www.google.com/enterprise

Gulli, A. & Signorini A. (2005). Posters: The indexable web is more than 11.5 billion pages. *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web WWW '05,* ACM, New York, NY.

Lyman, P. & Varian, H.R. (2003). How much information 2003? Retrieved from www.sims.berkeley.edu/research/projects/how-much-info-2003/

MPF. (2006). *The Millennium Technology Prize 2006*. Millennium Prize Foundation, Helsinki, Finland.

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. (2004). A heuristic information retrieval study: An investigation of methods for enhanced searching of distributed data objects exploiting bidirectional relevance feedback. PhD thesis, University of Bedfordshire.

Petratos, P. (2006). Information retrieval systems: A perspective on human computer interaction. *Journal of Issues in Informing Science and Information Technology*, *3*, 511-518. Available at http://informingscience.org/proceedings/InSITE2006/IISITPetr231.pdf

Petratos, P., Chen, L., Wang, P. & 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.

Petratos, P. & Gleni, S. (2006). Disinformation methods of financial crime via email, *Proceedings of the International Conference on Information Quality*, Massachusetts Institute of Technology, Boston, Massachusetts.

Porter, M.F. (1997). An algorithm for suffix stripping. In K. Sparck Jones & P. Willett (Eds.), *Readings in information retrieval*. San Francisco: Morgan Kaufmann.

Salton, G. (1989). *Automatic text processing.* Reading, Massachusetts: Addison Wesley.

Salton, G. & McGill, M.J. (1997). The SMART and SIRE experimental retrieval systems. In K. Sparck Jones & P. Willett (Eds.), *Readings in information retrieval.* San Francisco: Morgan Kaufmann.

Staikos, K. (2000). *The great libraries: From antiquity to the Renaissance, 3000 B.C. to A.D.1600*. New Castle, Delaware: Oak Knoll Press.

Varian, H.R. (2005). The digital society: Universal access to information. *Communications of the ACM, 48*(10).

Biography

Panagiotis Petratos is Assistant Professor of Computer Information Systems at California State University, Stanislaus. His research interests include information retrieval systems, human computer interaction, networking, computer security and biometrics enabled computer systems.