Information Retrieval Systems: A Perspective on Human Computer Interaction

Panagiotis Petratos
California State University, Stanislaus, Turlock, CA, USA

ppetratos@csustan.edu

Abstract

Traditional information systems design and development methodologies tend to overly focus on the technical details of the system such as memory management, system internals, algorithms and modules. It is not unusual for system designers and developers to often completely omit from the thought process the human element. This article offers a new information systems perspective particularly for information retrieval systems with a focus on human computer interaction.

Keywords: Information retrieval, human computer interaction, relevance feedback

Introduction

During the last few decades the amount of digital information has been constantly increasing at an accelerating rate. This condition is often referred to as the information overload problem. The reasons for this information overload problem are of technological as well as of social and economic nature.

Technological reasons include the recent advances in high density information storage as well as high speed telecommunications. Currently there are four types of media for the storage of information, magnetic, optical, film and the good old paper. Analogously, at present there are four types of streams for the dissemination of electronic information, the Internet, telephone, television and radio. A more traditional non-electronic but still very powerful source of information is the formal printed press which includes periodic publications such as newspapers and journals and non-periodic publications such as books.

It is interesting to note that the majority of new information created during 2002, or more than 98% of the total of all information transmitted in electronic information flows was exchanged between individuals during communication sessions including email, peer to peer file exchanges, and the telephone including both voice and data such as fax on land lines and instant messaging on wireless mobiles (Lyman & Varian, 2003).

Currently in the well-developed countries found in the continents of North America and Western Europe there is a digital convergence trend evident. Telecommunications companies such as Vonage are currently attempting to offer a flat fee model by diverting the enormous amount of information from telephone communications through the Internet and routing it as Internet packet traffic primarily due to economic reasons as the per minute charges are eliminated. It is only a matter of time before the Internet becomes

the primary backbone for telephone communications and the principal information flow infrastructure for all information transmitted.

Information Storage Technologies

Furthermore, recent magnetic storage technology advances are offering an alternative to the rapidly approaching physical limit of magnetic disk storage density. During the last fifty years the traditional storing process for magnetic disks has been utilizing longitudinal recording technology. The data bits of hard disks are microscopic magnetic grains that are aligned horizontally on the surface of the disk and are magnetized according to the longitudinal recording; please see Figure 1.

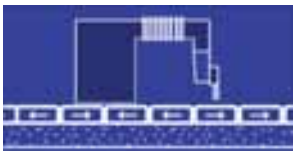

Figure 1: Longitudinal recording (Grochowski, 2005).

Based on this technology there is only one way to increase the data density of hard disks. This approach has been used successfully for the last fifty years steadily increasing data density by continuous miniaturization of the data bits. However, there are practical as well as physical limits to miniaturization, just as it is evident now to the microprocessor industry where the strategy of manufacturing ever smaller transistors, in order to pack as many as possible inside a single chip, is beginning to encounter its physical limits.

For the hard disk industry the practical as well as the physical limits are elucidated by a natural phenomenon called superparamagnetism, which occurs when the microscopic magnetic grains on the disk surface become so minute that haphazard thermal vibrations at room temperature are the aetiology for them to lose their ability to hold their magnetic orientations. The corollary is instant damage to data integrity that occurs as the microscopic magnetic grains whose north and south poles suddenly and arbitrarily reverse and experience magnetic anastrophe leading to data corruption and storage device unreliability.

A different approach from the existing data storing strategies is the new perpendicular recording technology; please see Figure 2. Perpendicular recording utilizes the depth of the disk, not just the surface of the circular platter, by aligning the microscopic magnetic grains vertically and thus allows for more data bits to be stored in the same space, which under longitudinal recording it was occupied by much fewer horizontal data bits (Mallary, Torabi, & Benakli, 2002). Perpendicular recording not only enables higher data densities but also overcomes the problem presented by the superparamagnetic effect.

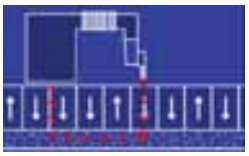

Figure 2: Perpendicular recording (Grochowski, 2005).

In March 2005, a data density of 230 gigabits per square inch (Gb/in^2) was demonstrated by Hitachi Global Storage Technologies utilizing the new perpendicular recording technology (Grochowski, 2005). This is the highest data density achieved heretofore based on perpendicular recording. This information storage accomplishment represents a twofold increase of the contemporary highest data densities in existence, which currently utilize longitudinal recording technology.

All these recent technological advances such as perpendicular recording and high speed internet access which enable people to store and disseminate increased amounts of information and the contemporary social trends which reveal that users are spending most of their time with electronic communications and exchange of information, they all indicate that information will continue increasing and more efficient systems for the search, organization, management and retrieval of information will be required.

Information Retrieval Systems

Information retrieval research and development over the years have produced a plethora 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 as shown in Figure 3 (Paice, 1990; Van Rijsbergen, 1979).

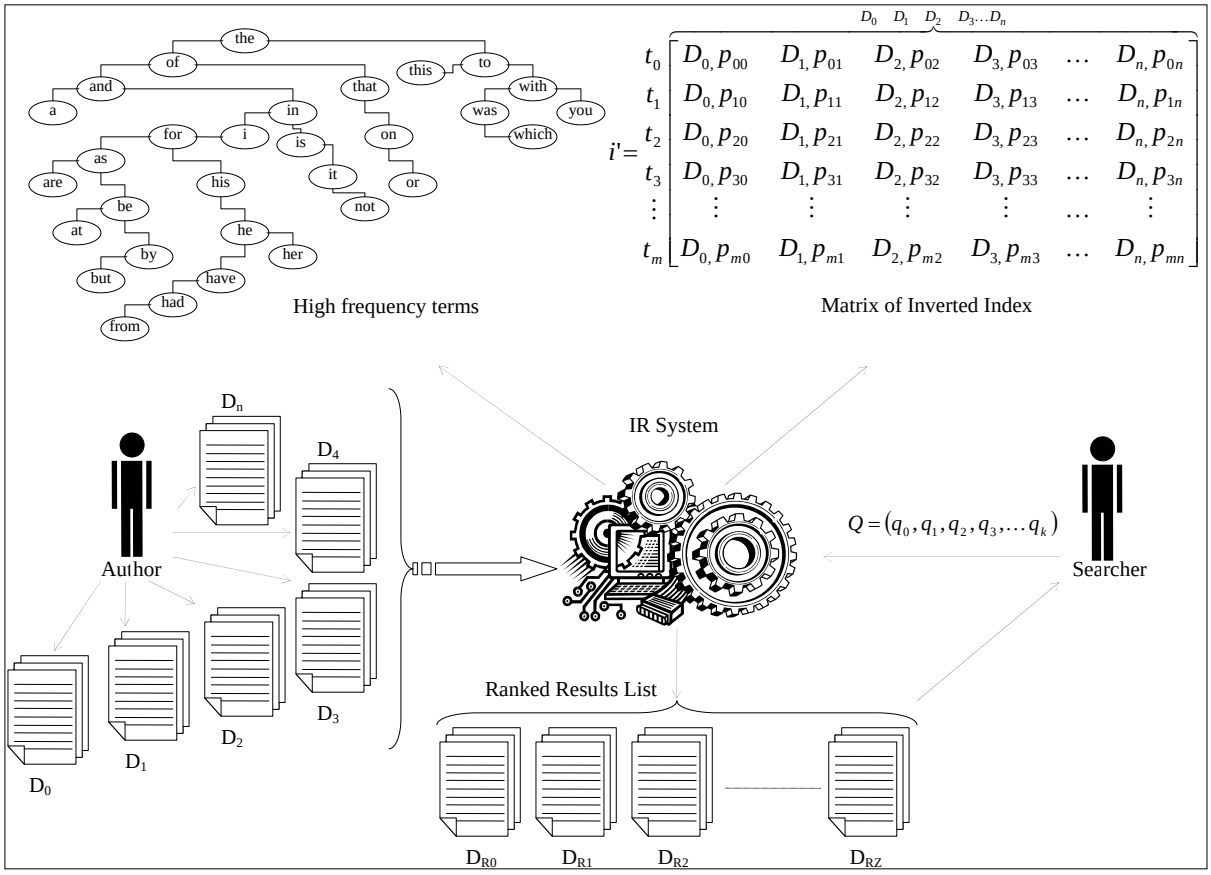

Figure 3: Traditional information retrieval system.

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 (Mauldin, 1991; Rillof & Lehnert, 1994).

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. For instance, look after means take care of, look back on means remember nostalgically, look down on means feel superior to others, look forward to means anticipate, look up to means respect and admire, look at a painting, look familiar, look happy.

If a look 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, man hunts shark, or an alleged murderer is not a murderer until proven guilty, or Lily has a beautiful Lotus automobile, or Jane slept away from lethal danger. In the first instance the information retrieval system returns an equivalency result that ensues if a query of a shark hunts man is issued.

In the second instance the information retrieval system returns an equivalency result which ensues if a query of a guilty murderer is issued. In the third instance the information retrieval system returns an equivalency result that ensues if a query of Lotus water Lily species is issued. In the third instance if a sleep disorder 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 et al., 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 anthropologic ability to read and understand.

Information retrieval experimental data analysis

In order to carry out a series of experiments an appropriate environment must be in place. Hence the development of the experimental information retrieval system *ANACALYPSE.* The architectonic model of the experimental information retrieval system *ANACALYPSE* is designed according to distributed object oriented architecture. A plethora of components are esoteric among didymous exoteric entities that coordinate the interactions and communication of their esoteric collections of objects. The didymous distinct exoteric entities are the server and the client(s), which can be arranged to operate in a parallel configuration in order for a plethora of simultaneous clients to interact synchronously in parallel with a monadic server.

Each client consists of a plethora of objects such as interactive dialogue and user input components, user historical data preservation modules, 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 plethora of modules for various operations. A few of these operations include the following:

> Locating and transferring from distributed machines a plethora of relevant documents 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 photographs, acoustic and cinematographic files. Identifying and extracting undesired high frequency terms. Discovering and preserving the types and roots of words.
>
> Esoterically preserving the machine's historical data. Performing the analysis and metamorphosis of 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.
>
> Automatically performing the syntheses of experts' bidirectional fuzzy relevance feedback with the machine training sample of corresponding document vectors in a new *Metagramma* matrix vector structure. Automatically computing the term weights of the *Metagramma* matrix vector structure according to the experts' bidirectional fuzzy relevance feedback on the machine training sample of documents.
>
> Automatically computing the similarity of the *Metagramma* matrix vector structure with all the term vectors of all the documents that 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.

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 experts' 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 ergo the group impact for a particular system can then be precisely computed. The expert supervised training of the bidirectional fuzzy information retrieval system can be conducted in order to eventually achieve unsupervised system document relevance classification based on the initial small training sample and a relevance function which is morphed through syntheses of the supervised training, the document similarity measure and the term weighting function.

Dyads of supervised training modes are possible such as unidirectional and bidirectional fuzzy relevance supervised training. Firstly, the standard unidirectional fuzzy interval is [0, 1] and the documents which score 0 are completely irrelevant, the documents that score ½ are neutral and the documents that score 1 are highly relevant. Secondly, 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. The queries are selected in the area of expertise by the expert supervisor in order for the relevance training to be accurate and a few of the queries selected by the expert supervisors are for instance “Itanium instruction set”, “California commercial insurance laws”, “Itanium source code porting methodologies”, “Artificial intelligence algorithms”, “Fuzzy logic techniques”, “Data mining methodologies”, etc. 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 present study summarizes the results obtained with the *ANACALYPSE* information retrieval system based on the processing of *2,000* documents and presents evaluation output in comparison to Google a commercial information retrieval system as shown next.

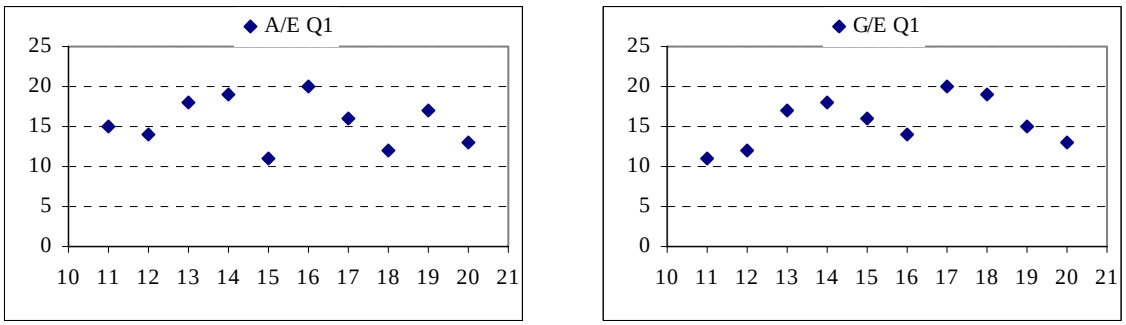

Figure 4. Scatter plots for q1 Anacalypse/Expert (left), Google/Expert (right) correlations.

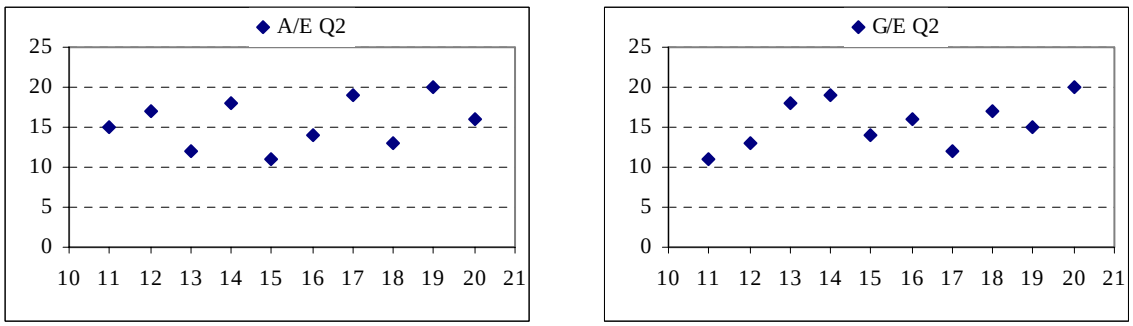

Figure 5. Scatter plots for q2 Anacalypse/Expert (left), Google/Expert (right) correlations.

In synopsis, the primary scope of this 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:

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 perform a triadic comparison and evaluate the effectiveness of the experimental information retrieval 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 is conducted in order to eventually achieve unsupervised system document relevance classification based on an initial small training sample and a relevance function which is morphed 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.

The procedure is based on a set of queries, which are selected in the research area (i.e. artificial intelligence, bioinformatics, software engineering, information retrieval, decipherment, etc.) of expertise by the experts who are researchers at the University of Luton in order for the relevance training and Metagramma 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 20 queries a grand total of 2,000 documents were retrieved and processed by *ANACALYPSE* for which the average Spearman's correlation to the experts' relevance judgments was 0.46542 whilst for Google the average Spearman's correlation to the experts' relevance judgments was 0.256995. According to these experimental results *ANACALYPSE*, the experimental information retrieval system is closer to the expert's relevance ranks with an average 10% positive change over Google, the commercial information retrieval system.

Conclusion

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 that represent additional effort in time and reading for the information seeker. In this study syntheses of bidirectional fuzzy logic and information retrieval methodologies have been developed, analyzed and evaluated by subject matter experts. The novel methodologies have been designed and implemented in an experimental information retrieval system. According to the experimental results the academic information retrieval system is closer to the expert's relevance judgments with an average 10% positive auxesis over Google, the commer-cial information retrieval system. The proposed methodologies elucidate the bases for improved effectiveness and efficiency of information retrieval.

References

Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. & Harshman, R.A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Grochowski, E. (2005). *Hard disk technology overview*. Hitachi Global Storage Technologies, San Jose Research Center. Retrieved January 9, 2006, from http://www.hitachigst.com/hdd/hddpdf/tech/hdd_technology2003.pdf

Hudson, R. (1995). *Word Meaning.* London, England: Routledge.

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*, 37-50. New York, New York: ACM.

Liddy, E., Paik, W. & 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.

Lyman, P. & Varian, H. R. (2003). *How much information*? UC Berkeley School of Information Management & Systems. Retrieved January 9, 2006, from http://www.sims.berkeley.edu/research/projects/how-much-info-2003

Mallary, M., Torabi, A., & Benakli, M (2002). One terabit per square inch perpendicular recording conceptual design. *IEEE Transactions on Magnetics*, *38*(4 I), 1719-1724.

Mauldin, M.L. (1991). *Conceptual information retrieval – A case study in adaptive partial parsing*. Boston, Massachusetts: Kluwer Academic Publishers.

Paice, C. (1990). Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management, 26*(1), 171-186.

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.

Van Rijsbergen, C.J. (1979). *Information Retrieval* (2nd ed.). London, England: Butterworths.

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*, 454-460, Nantes, France.

Biography

Panagiotis Petratos is Assistant Professor of Computer Information Systems at California State University, Stanislaus. His research interests include object oriented system design, information retrieval systems, database systems, computer security, biometrics enabled computer systems. Email: ppetratos@csustan.edu