PERCEPTION-BASED INFORMATION PROCESSING

by

Masoud Nikravesh and Dae-Young Choi

Memorandum No. UCB/ERL M03/20

10 June 2003

PERCEPTION-BASED INFORMATION PROCESSING

by

Masoud Nikravesh and Dae-Young Choi

Memorandum No. UCB/ERL M03/20

10 June 2003

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Perception-Based Information Processing

Masoud Nikravesh [1] and Dae-Young Choi[1,2]

[1] BISC Program, EECS Department-CS Division
University of California, Berkeley, CA 94720
Nikravesh@cs.berkeley.edu
[2] Dept. of MIS, Yuhan College, Koean-Dong, Sosa-Ku,
Puchon City, Kyungki-Do, South Korea
dychoi@green.yuhan.ac.kr

Abstract: Humans have a remarkable capability (perception) to perform a wide variety of physical and mental tasks without any measurements or computations. Familiar examples of such tasks are: playing golf, assessing wine, recognizing distorted speech, and summarizing a story. The question is whether a special type information retrieval processing strategy can be designed that build in perception. Commercial Web search engines have been defined which manage information only in a crisp way. Their query languages do not allow the expression of preferences or vagueness. Even though techniques exist for locating exact matches, finding relevant partial matches might be a problem. It may not be also easy to specify query requests precisely and completely - resulting in a situation known as a fuzzy-querying. It is usually not a problem for small domains, but for large repositories such as World Wide Web, a request specification becomes a bottleneck. Thus, a flexible retrieval algorithm is required, allowing for imprecise or fuzzy query specification or search. In addition, they have problems as follows : (1) large answer set ; (2) low precision; (3) unable to preserve the hypertext structures of matching hyperdocuments; (4) ineffective for general-concept queries. The task is to use user-defined queries to retrieve useful information according to certain measures. In order to handle these problems, we propose the Perception Index (PI) that contains attributes associated with a focal keyword restricted by fuzzy term(s) used in fuzzy queries on the Internet. If we integrate the Document Index (DI) used in commercial Web search engines with the proposed PI, we can handle both crisp terms (keyword-based) and fuzzy terms (perception-based). In this respect, the proposed approach is softer than the keyword-based approach. The PI brings somewhat closer to natural language. It is a further step toward a real human-friendly, natural language-based interface for Internet. It should greatly help the user relatively easily retrieve relevant information. In other words, the PI assists the user to reflect his/her perception in the process of query. Conse-

quently, Internet users can narrow thousands of hits to the few that users really want. In this respect, the PI provides a new tool for targeting queries that users really want, and an invaluable personalized search. In this chapter, we also present the search mechanism based on the integrated index (DI + PI) and fuzzy query based on the integrated index (DI + PI). Moreover, we describe some features of the proposed method and suggest some considerations for implementing the proposed method. The main goal of the perception-based information processes and retrieval system is to design a model for the internet based on user profile with capability of exchanging and updating the rules dynamically and *"do what I mean, not as I say"* and using programming with *"human common sense capability"*.

1 Introduction

Under leadership of DARPA, ARPANET has been designed through close collaboration with UCLA during 1962-1969, 1970-1973, and 1974-1981. Initially designed to keep military sites in communication across the US. In 1969, ARPANET connected researchers from Stanford University, UCLA, UC Santa Barbara and the University of Utah. The Internet community formed in 1972 and the Email is started in 1977. While initially a technology designed primarily for needs of the U.S. military, the Internet grew to serve the academic and research communities. More recently, there has been tremendous expansion of the network both internationally and into the commercial user domain.

There are many publicly available Web search engines, but users are not necessarily satisfied with speed of retrieval (i.e., slow access) and quality of retrieved information (i.e., inability to find relevant information). It is important to remember that problems related to speed and access time may not be resolved by considering Web information access and retrieval as an isolated scientific problem. An August 1998 survey by Alexa Internet (<alexa.com>) indicates that 90% of all Web traffic is spread over 100,000 different hosts, with 50% of all Web traffic headed towards the top 900 most popular sites. Effective means of managing uneven concentration of information packets on the Internet will be needed in addition to the development of fast access and retrieval algorithms (Kabayashi and Takeda 2000).

World Wide Web search engines have become the most heavily-used online services, with millions of searches performed each day. Their popularity is due, in part, to their ease of use. The central tasks for the most of the search engines can be summarize as 1) query or user information request- do what I mean and not what I say!, 2) model for the Internet, Web representation-web page collection, documents, text, images, music, etc, and 3) ranking or matching function-degree of relevance, recall, precision, similarity, etc. ***Table 1*** also compares the issues related to the conventional Database with Internet. Already explosive amount of

users on the Internet is estimated over 200 million (***Table 2***). While the number of pages available on the Internet almost double every year, the main issue will be the size of the internet when we include multimedia information as part of the Web and also when the databases connected to the pages to be considered as part of an integrated Internet and Intranet structure. Databases are now considered as backbone of most of the E-commerce and B2B and business and sharing information through Net between different databases (Internet-Based Distributed Database) both by user or clients are one of the main interest and trend in the future. In addition, the estimated user of wireless devices is estimated 1 billion within 2003 and 95 % of all wireless devices will be Internet enabled within 2005. ***Table 3*** shows the evolution of the Internet, World Wide Web, and Search Engines.

Table 1. Database Vs. Internet

Database	Internet
Distributed	Distributed
Controlled	Autonomous
Query (QL)	Browse (Search)
Precise	Fuzzy/Imprecise
Structure	Unstructured

Table 2. Internet and rate of changes

Jan 1998: 30 Millions web hosts
Jan 1999: 44 Millions web hosts
Jan 2000: 70 Millions web hosts
Feb 2000: +72 Millions web hosts

Dec 1997: 320 Millions
Feb 1999: 800 Millions
March 2000: +1,720 Millions

The number of pages available on the Internet almost doubles every year

Courtois and Berry (Martin P. Courtois and Michael W. Berry, ONLINE, May 1999-Copyright © Online Inc.) published a very interesting paper "Results Ranking in Web Search Engines". In their work for each search, the following topics were selected: credit card fraud, quantity theory of money, liberation tigers, evolutionary psychology, French and Indian war, classical Greek philosophy, Beowulf criticism, abstract expressionism, tilt up concrete, latent semantic indexing, fm synthesis, pyloric stenosis, and the first 20 and 100 items were downloaded using the search engine. Three criteria 1) All Terms, 2) Proximity, and 3) Location were used as a major for testing the relevancy ranking. ***Table 4*** shows the concept of relevancy and its relationship with precision and recall (***Table 5*** and ***Figure 1***). ***Table 6*** shows the summary of the results. The effectiveness of the classification is defined based on the precision and recall (***Tables 4-5*** and ***Figure 1***).

Table 4. Similarity/Precision and Recall

	Relevant	Non-Relevant	
Retrieved	$A \cap B$	$\bar{A} \cap B$	B
Not Retrieved	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$\bar{B}$
	A	$\bar{A}$	N

N: Number of documents

Table 5. Similarity/Measures of Association

There are five commonly used measures of association in IR :

Simple matching Coefficient : $|X \cap Y|$

Dice's Coefficient : $2\frac{|X \cap Y|}{|X| + |Y|}$

Jaccard's Coefficient : $\frac{|X \cap Y|}{|X \cup Y|}$

Cosine Coefficient : $\frac{|X \cap Y|}{|X|^{1/2} \times |Y|^{1/2}}$

Overlap Coefficient : $\frac{|X \cap Y|}{\min(|X|, |Y|)}$

Dissimilarity Coefficient : $\frac{|X \Delta Y|}{|X| + |Y|} = 1 - $ ***Dice's Coefficient***

$|X \Delta Y| = |X \cup Y| - |X \cap Y|$

Table 6. Results Ranking in Web Search Engines

Criteria	All Terms		Proximity		Location	
	20/100 hits	Mean hits	20/100 hits	Mean hits	20/100 hits	Mean hits
First 20 and 100 items						
ALTAVISTA	31/13%	22%	11/7%	9%	41/10%	25.5%
EXCITE	18/5%	11.5%	28/5%	16.5%	77/53%	65%
HOTBOT	19/12%	15.5%	40/24%	32%	62/29%	45.5%
INFOSEEK	23/16%	19.5%	14/10%	12%	79/50%	64.5%
LYCOS	8/5%	6.5%	49/26%	37.5%	69/32%	50.5%

Effectiveness is a measure of the system ability to satisfy the user in terms of the relevance of documents retrieved. In probability theory, precision is defined as conditional probability, as the probability that if a random document is classified under selected terms or category, this decision is correct. Precision is defined as portion of the retrieved documents that are relevant with respect to all retrieved documents; number of the relevant documents retrieved divided by all documents retrieved. Recall is defined as the conditional probability and as the probability if a random document should be classified under selected terms or category, this decision is taken. Recall is defined as portion of the relevant retrieved documents that are relevant with respect to all relevant documents exists; number of the relevant documents retrieved divided by all relevant documents. The performance of each request is usually given by precision-recall curve (***Figure 1***). The overall performance of a system is based on a series of query request. Therefore, the performance of a system is represented by a precision-recall curve, which is an average of the entire precision-recall curve for that set of query request.

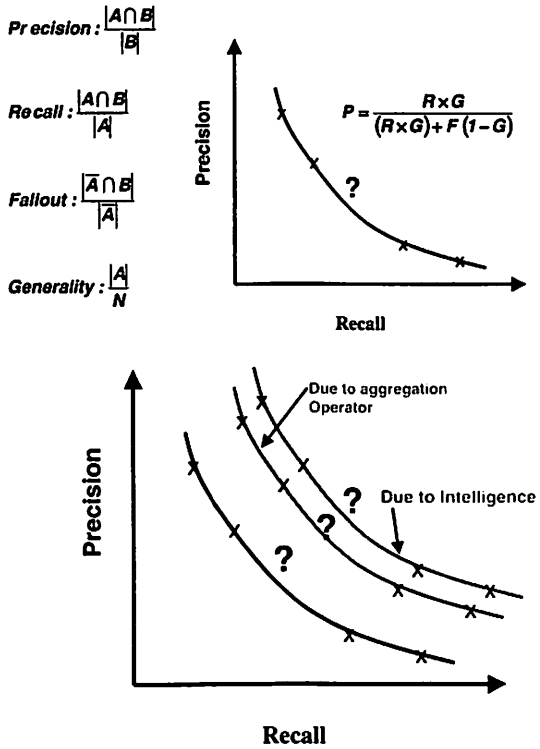


Figure 1. **1.a.)** relationship between Precision and Recall, **1.b.)** inverse relationship between Precision and Recall.

To improve the performance of a system one can use different mathematical model for aggregation operator for (A∩B) such as fuzzy logic. This will sift the curve to a higher value as is shown in ***Figure 1.b***. However, this may be a matter of scale change and may not change the actual performance of the system. We call this improvement, virtual improvement. However, one can shit the curve to the next level, by using a more intelligent model that for example have deductive capability or may resolve the ambiguity (***Figure 1.b***).

Many search engines support Boolean operators, field searching, and other advanced techniques such as fuzzy logic in variety of definition and in a very primitive ways (***Table 7***). While searches may retrieve thousands of hits, finding relevant partial matches and query relevant information with deductive capabilities might be a problem. ***Figure 2.*** shows a schematic diagram of model presented by Lotfi A. Zadeh (2002) for the flow of information and decision. What is also important to mention for search engines is query-relevant information rather than generic information. Therefore, the query needs to be refined to capture the user's perception. However, to design such a system is not trivial, however, Q/A systems information can be used as a first step to build a knowledge based to capture some of the common user's perceptions. Given the concept of the perception, new machineries and tools need to be developed. Therefore, we envision that non-classical techniques such as fuzzy logic based-clustering methodology based on perception, fuzzy similarity, fuzzy aggregation, and FLSI for automatic information retrieval and search with partial matches are required.

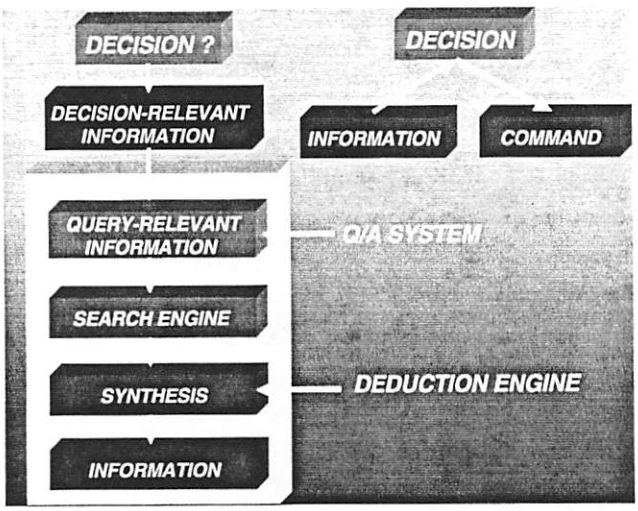


Figure 2. Perception-Based Decision Analysis (PDA) (Zadeh, 2001)

Table 7. Examples of Fuzzy Web Search Engines

Search Engine	Simple Form	Search Logic				Fuzzy Logic in any form	Term Weighting	Sorted Output	Ranked output	Find Like
		Boolean	Proximity	Nesting	Truncation					
Excite!	X	X	X	X		X	X	X	X	X
AltaVista	X	X	X	X	X			X	X	
HotBot		X	X	X		X	X		X	
Infoseek	X	X	X	X	X	X	X	X	X	
Lycos	X				X	X*			X	
Open Text		X	X	X					X	(X)
Web Crawler	X	X	X	X		X	X	X	X	
Yahoo	X	X	X	X			X			
Google	X	X	*	*	*	X	*	*	*	*
Northern Light	X	X	*	*	*	X	*	*	*	*
Power Fast Search Advanced	X	X	*	*	*	X	*	*	*	*

2 Intelligent Search Engines

Design of any new intelligent search engine should be at least based on two main motivations:

- The web environment is, for the most part, unstructured and imprecise. To deal with information in the web environment what is needed is a logic that supports modes of reasoning which are approximate rather than exact. While searches may retrieve thousands of hits, finding decision-relevant and query-relevant information in an imprecise environment is a challenging problem, which has to be addressed.
- Another, and less obvious, is deduction in an unstructured and imprecise environment given the huge stream of complex information.

Tim Bernres-Lee (1999) in his transcript refers to the fuzzy concept and the human intuition with respect to the Web (Transcript of Tim Berners-Lee's talk to the LCS 35th Anniversary celebrations, Cambridge Massachusetts, 1999/April/14):

Lotfi A. Zadeh (2001a) consider fuzzy logic is a necessity to add deductive capability to a search engine: "Unlike classical logic, fuzzy logic is concerned, in the main, with modes of reasoning which are approximate rather than exact. In Internet, almost everything, especially in the realm of search, is approximate in nature. Putting these two facts together, an intriguing thought merges; in time, fuzzy logic may replace classical logic as what may be called the brainware of the Internet.

...

In my view, among the many ways in which fuzzy logic may be employed, there are two that stand out in importance. The first is search. Another, and less obvious, is deduction in an unstructured and imprecise environment. Existing search engines have zero deductive capability. ... To add a deductive capability to a search engine, the use of fuzzy logic is not an option - it is a necessity."

With respect to the deduction and its complexity, Lotfi's viewpoint (2001a and 2002) is summarized as follows: "Existing search engines have many remarkable capabilities. But what is not among them, is the deduction capability -- the capability to answer a query by drawing on information which resides in various parts of the knowledge base or is augmented by the user. Limited progress is achievable through application of methods based on bivalent logic and standard probability theory. But to move beyond the reach of standard methods it is necessary to

change direction. In the approach, which is outlined, a concept which plays a pivotal role is that of a prototype -- a concept which has a position of centrality in human reasoning, recognition, search and decision processes. ... The concept of a prototype is intrinsically fuzzy. For this reason, the prototype-centered approach to deduction is based on fuzzy logic and perception-based theory of probabilistic reasoning, rather than on bivalent logic and standard probability theory. What should be underscored, is that the problem of adding deduction capability to search engines is many-faceted and complex. It would be unrealistic to expect rapid progress toward its solution."

During 80, most of the advances of the automatic document categorization and IR were based on knowledge engineering. The models were built manually using expert systems capable of taking decision. Such expert system has been typically built based on a set of manually defined rules. However, the bottleneck for such manual expert systems was the knowledge acquisition very similar to expert system. Mainly, rules needed to be defined manually by expert and were static. Therefore, once the database has been changed or updated the model must intervene again or work has to be repeated anew if the system to be ported to a completely different domain. By explosion of the Internet, these bottlenecks are more obvious today. During 90, new direction has been merged based on machine learning approach. The advantage of this new approach is evident compared to the previous approach during 80. In machine learning approach, most of the engineering efforts goes towards the construction of the system and mostly is independent of the domain. Therefore, it is much easier to port the system into a new domain. Once the system or model is ported into a new domain, all that is needed is the inductive, and updating of the system from a different set of new dataset, with no required intervention of the domain expert or the knowledge engineer. In term of the effectiveness, IR techniques based on machine learning techniques achieved impressive level of the performance and for example made it possible automatic document classification, categorization, and filtering and making these processes viable alternative to manual and expert system models.

Doug B. Lenat both the founder of the CYC project and president of Cycorp (http://www.cyc.com) puts the concept of deduction into perspective and he expresses that both commonsense knowledge and reasoning are key for better information extraction (2001).

Lotfi A. Zadeh (2002) express qualitative approach towards adding deduction capability to the search engine based on the concept and framework of protoforms:

"At a specified level of abstraction, propositions are p-equivalent if they have identical protoforms." "The importance of the concepts of protoform and p-equivalence derives in large measure from the fact that they serve as a basis for knowledge compression."

"A knowledge base is assumed to consist of a factual database, FDB, and a deduction database, DDB. Most of the knowledge in both FDB and DDB is per-

ception-based. Such knowledge cannot be dealt with through the use of bivalent logic and standard probability theory. The deduction database is assumed to consist of a logical database and a computational database, with the rules of deduction having the structure of protoforms. An example of a computational rule is "if Q_1 A's are B's and Q_1 (A and B)'s are C's," then "$Q_1 Q_2$ A's are(B and C)'s, where Q_1 and Q_2 are fuzzy quantifiers and A, B and C are labels of fuzzy sets. The number of rules in the computational database is assumed to be very large in order to allow a chaining of rules that may be query-relevant."

Computational theory of perception (CTP) (Zadeh, 1999 and 2001b; Nikravesh et al., 2001; Nikravesh, 2001a and 2001b) is one of the many ways that may help to address some of the issues presented by both Berners Lee and Lotfi A. Zadeh earlier, a theory which comprises a conceptual framework and a methodology for computing and reasoning with perceptions. The base for CTP is the methodology of computing with words (CW) (Zadeh 1999). In CW, the objects of computation are words and propositions drawn from a natural language.

3 Perception-Based Information Processing for Internet

One of the problems that Internet users are facing today is to find the desired information correctly and effectively in an environment that the available information, the repositories of information, indexing, and tools are all dynamic. Even though some tools were developed for a dynamic environment, they are suffering from "too much" or " too little" information retrieval. Some tools return too few resources and some tool returns too many resources *(Figure 3).*

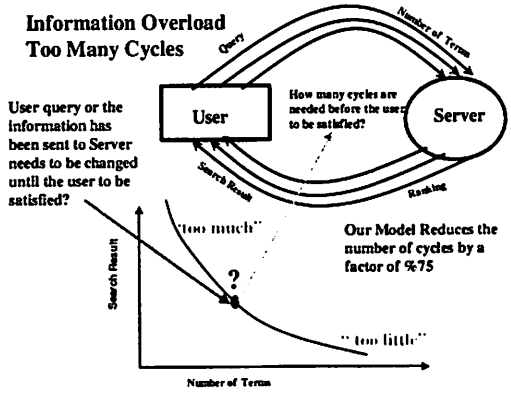


Figure 3. Information overload

The main problem with conventional information retrieval and search such as vector space representation of term-document vectors are that 1) there is no real theoretical basis for the assumption of a term and document space and 2) terms and documents are not really orthogonal dimensions. These techniques are used more for visualization and most similarity measures work about the same regardless of model. In addition, terms are not independent of all other terms. With regards to probabilistic models, important indicators of relevance may not be term -- though terms only are usually used. Regarding Boolean model, complex query syntax is often misunderstood and problems of null output and Information overload exist. One solution to these problems is to use extended Boolean model or fuzzy logic. In this case, one can add a fuzzy quantifier to each term or concept. In addition, one can interpret the AND as fuzzy-MIN and OR as fuzzy-MAX functions. Alternatively, one can add agents in the user interface and assign certain tasks to them or use machine learning to learn user behavior or preferences to improve performance. This technique is useful when past behavior is a useful predictor of the future and wide variety of behaviors amongst users exist.

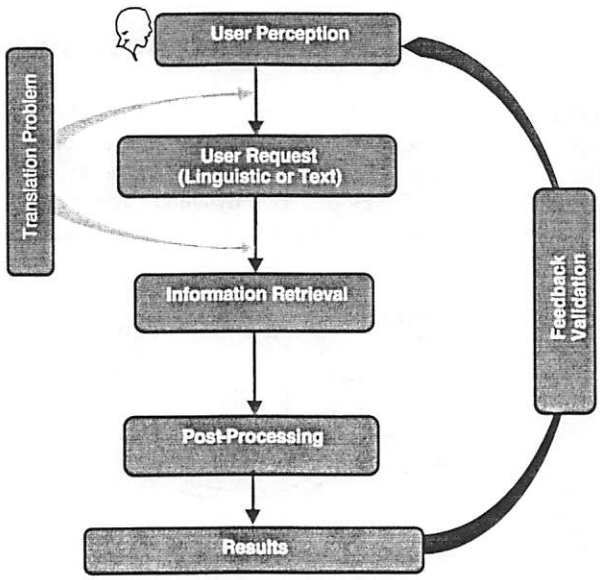


Figure 4.a. Structure of conventional search engine and retrieval technique

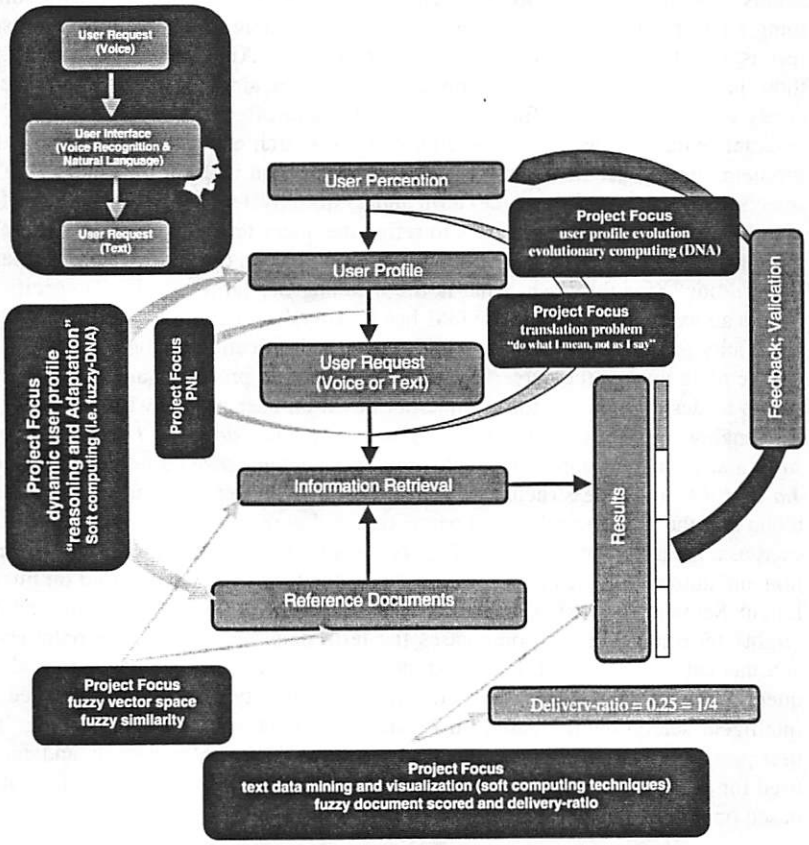


Figure 4. b Structure of search engine and retrieval technique and the problem related to perception and areas that soft computing can be used as a mean for improvement.

In addition, the user's perception, which is one of the most important key features, is oftentimes ignored. For example, consider the word " football". The perception of an American differs from the perception of an European who understands football to mean "Soccer." Therefore, if the search engine knows something about the user and its perception, it might be able to better refine the users results. For this example, there is no need to eliminate American football pages for those in the UK looking for real football information, since this information inclusively exists in user's profile. Search Engines also often return a large list of irrelevant search results due to the ambiguity of search query terms. To solve this problem one can use the following approaches 1) from Users Side/ Client Side by selecting a very specific (unique) term and 2) from Systems Sides/Server by offering alternate query terms for users to refine the query terms. Sources of the ambiguity are mainly due to 1) definition/meaning and as an example-what is the largest building? (for this case, what is the meaning of "largest") and 2) specificity and as an example- where is the GM headquarters? (for this case, what level of specificity is required?). To address this issue, a clarification dialog is required.

The main goal of the perception-based information processes and retrieval system is to design a model for the internet based on user profile with capability of exchanging and updating the rules dynamically and *"do what I mean, not as I say"* and using programming with *"human common sense capability"*. ***Figures 4.a and 4.b*** show the structure of conventional search engine and retrieval technique and the problem related to perception and areas that soft computing can be used as a mean for improvement. ***Figure 5*** shows the automated ontology generation and automated document indexing using the terms similarity based on Fuzzy-Latent Semantic Indexing Technique (FLSI). Often time it is hard to find the "right" term and even in some cases the term does not exist. The ontology is automatically constructed from text document collection and can be used for query refinement. ***Figure 6*** shows documents similarity map that can be used for intelligent search engine based on FLSI, personalization and user profiling. The user profile is automatically constructed from text document collection and can be used for query refinement and provide suggestions and for ranking the information based on pre-existence user profile.

4 Fuzzy Conceptual Model and Search Engine

One can use clarification dialog, user profile, context, and ontology, into a integrated frame work to address some of the issues related to search engines were described earlier. In our perspective, we define this framework as ***Fuzzy Conceptual Matching based on Human Mental Model (Figure 7)***. The Conceptual Fuzzy Set (CFS) model will be used for intelligent information and knowledge retrieval through conceptual matching of both text and images (here defined as "Concept").

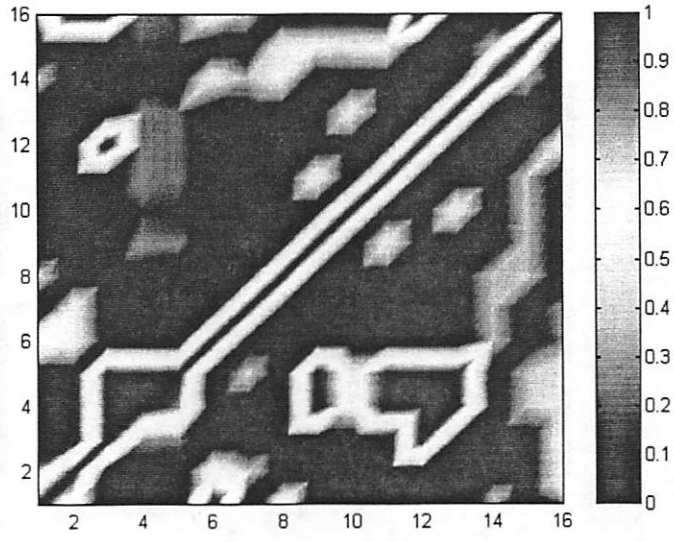


Figure 5. Terms Similarity; Automated Ontology Generation and Automated Indexing

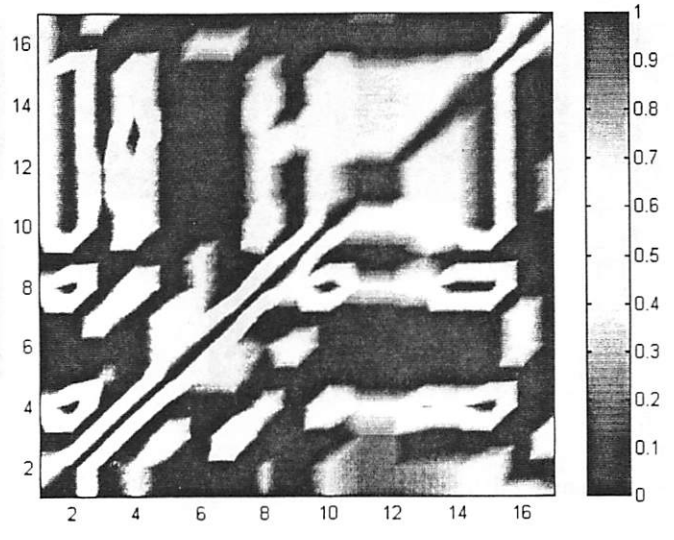


Figure 6. Documents Similarity; Search Personalization-User Profiling

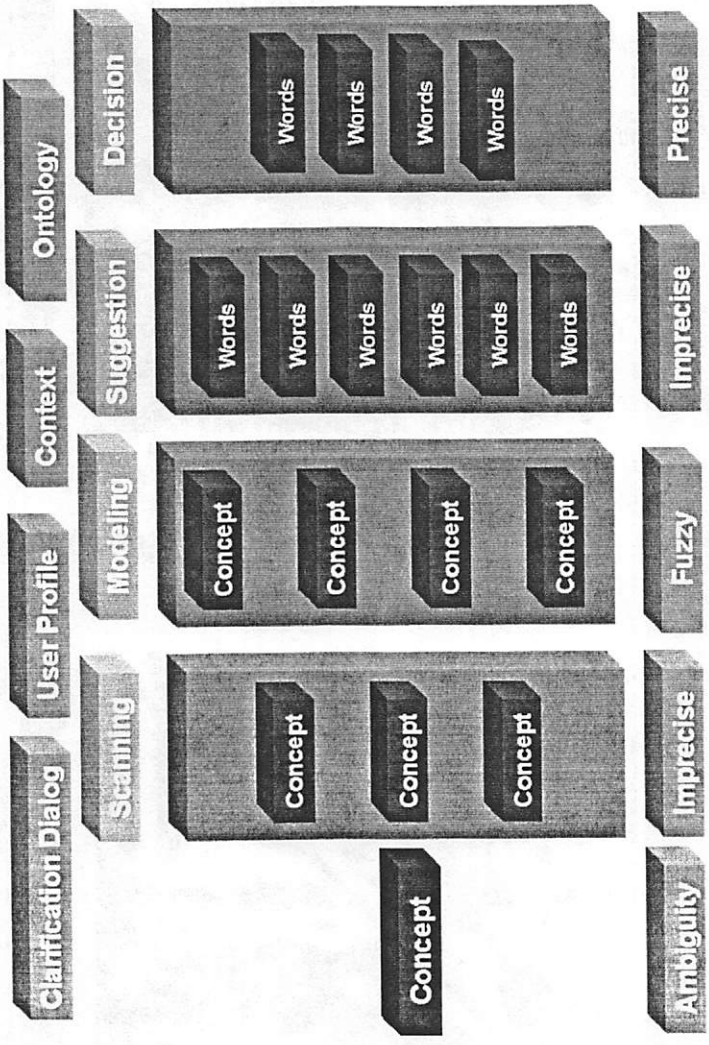


Figure 7. Fuzzy Conceptual Matching and Human Mental Model

The selected query doesn't need to match the decision criteria exactly, which gives the system a more human-like behavior. The CFS can also be used for constructing fuzzy ontology or terms related to the context of search or query to resolve the ambiguity. It is intended to combine the expert knowledge with soft computing tool. Expert knowledge needs to be partially converted into artificial intelligence that can better handle the huge information stream. In addition, sophisticated management work-flow need to be designed to make optimal use of this information. The new model can execute conceptual matching dealing with context-dependent word ambiguity and produce results in a format that permits the user to interact dynamically to customize and personalized its search strategy.

5 Fuzzy Query on the Internet

In this section, we do not attempt to solve 'unable to preserve the hypertext structures of matching hyperdocuments' problem. However, we try to tackle in part the other problems (i.e., 'large answer set', 'low precision', 'ineffective for general-concept queries'). In order to handle these problems, we propose the Perception Index (PI) that contains attributes associated with a focal keyword restricted by fuzzy term(s) used in fuzzy queries on the Internet.

5.1 Integration of Document Index with Perception Index

The central concept of information retrieval is the notion of relevance (Salton 1989). A user with a given query for information tries to find any specific results that he/she really wants. There are several models for specifying the representations used for the documents and the queries, as well as the matching of these representations (Kraft and Petry 1997). The most used model is that of the Boolean query based on set theory. Documents are represented as sets of terms and queries are Boolean expressions on terms. The retrieval mechanism does an exact match by classifying documents that satisfy the Boolean query as being relevant, all other documents as being irrelevant. This model is used by virtually all commercial textual-document retrieval systems. However, it is difficult to overcome the limitations of this model, including the inability to handle properly imprecision and subjectivity. The second model is the vector space model (Salton 1989) where documents and queries are represented as vectors in the space of all possible index terms. The document vectors consist of weights based on term frequencies in the collection, while the query vectors are binary vectors on the terms. The matching is based on a similarity measure between the documents and the query (often involving the cosine of the angle between the query vector and a given document vector). To date, this model leads the others in terms of performance. The third model is the probabilistic model (Salton 1989) where documents are represented as binary vectors. The queries are vectors of terms with weights based on the es-

timated probability of relevance of documents with those terms. Like the vector space model, the key advantage is the ability to rank documents on the likelihood of relevance. The fourth model is the generalized Boolean model, where fuzzy set theory allows the extension of the classical Boolean model to incorporate weights and partial matches, and adding the idea of document ranking.

The importance of representations of uncertainty in databases is increasing as more complex applications such as CAD/CAM and geographical information systems (GIS) are being undertaken in object-oriented and multi-media databases. Query languages are designed to express the user's retrieval requests in either a crisp manner or not. Much of the work in the database area has been in extending query languages to permit the representation and retrieval of imprecise data (Kacprzyk and Ziolkowski 1986, Nakajima et el. 1993, Petry and Bosc 1996, Rasmussen and Yager 1999, and Testemale 1986). There are some current commercial attempts at providing fuzzy query capabilities as front ends to conventional database systems (Nakajima et el. 1993).

Until now, however, commercial systems including informational retrieval systems (IRS), data base management systems (DBMS), and Web search engines have been defined which manage information only in a crisp way. Moreover, (crisp) traditional query languages do not allow the expression of preferences or vagueness which could be desirable for the following reasons (Kraft and Petry 1997):

- to control the size of the results;
- to express soft retrieval conditions;
- to produce a discriminated answer.

Although the commercial Web search engines such as Yahoo !, Google, Lycos, etc. help Internet users get to good information, they do not properly handle fuzzy query and tend to ignore the importance of fuzzy terms in a query. The expressive power of conventional search engine query interfaces is relatively weak when restricted to keyword-based search (i.e., Document Index (DI)-based search) (Kao et al. 2000).

In Section 5.2, we introduce the integrated index (DI + PI) and suggest a new search mechanism based on the integrated index. In Section 5.3, we describe fuzzy query based on the integrated index. This section is divided into three parts : types of fuzzy query, query processing based on the integrated index, and user interface based on the integrated index. In Section 5.4, we summarize some features of the proposed method. In Section 5.5, we show the effectiveness of our approach. In Section 5.6, we suggest some considerations for implementing the proposed method. We discuss the proposed method in Section 5.7.

5.2 Integration of Document Index with Perception Index

The most important of the tools for information retrieval is the index – a collection of terms with pointers to places where information about documents can be found. The development of effective indexing tools to aid in filtering is one of major classes of problems associated with Web search and retrieval. Removal of spurious information is a particularly challenging problem (Kobayashi and Takeda 2000).

Search engines are the most popular tools that people use to locate information on the Web. A search engine works by traversing the Web via the hyperlinks that connect the Web pages, performing text analysis on the pages it has encountered, and indexing the pages based on the keywords they contain. A user seeking information from the Web would formulate his/her information goal in terms of a few keywords composing a query. A search engine, on receiving a query, would match the query against its Document Index (DI). All of the pages that match the user query will be selected into an *answer set* and be ranked according to how relevant the pages are with respect to the query. Relevancy here is usually based on the number of matching keywords that a page contains (Kao et al. 2000). The DI is generally consisted of keywords that appear in the title of a page or in the text body. Based on the DI, the commercial Web search engines such as Yahoo !, Google, Lycos, etc. help users get to good information. For example, BigBook (or SuperPages) can help users to find 'Italian restaurants within a 1-mile radius from a specific address' (U.S. yellow pages services) (Lidsky and kwon 1997). This proximity search is processed based on crisp query with keywords (i.e., 'Italian restaurants', '1-mile', 'a specific address'). However, they do not properly process fuzzy queries. For example, find '***popular*** national parks in the USA'. In addition, they have problems as follows (Kao et al. 2000):

- large answer set;
- low precision;
- unable to preserve the hypertext structures of matching hyperdocuments;
- ineffective for general-concept queries.

In this section, we do not attempt to solve 'unable to preserve the hypertext structures of matching hyperdocuments' problem. However, we try to tackle in part the other problems (i.e., 'large answer set', 'low precision', 'ineffective for general-concept queries'). In order to handle these problems, we propose a Perception Index (PI). The remarkable human capability to perform a wide variety of physical and mental tasks without any measurements and any computations is derived from the brain's crucial ability to manipulate perceptions – perceptions of distance, size, weight, color, speed, time, direction, force, number, truth, likelihood, and other characteristics of physical and mental objects. Familiar examples

of the remarkable human capability are parking a car, driving in heavy traffic, playing golf, riding a bicycle, understanding speech, and summarizing a story (Zadeh 1999). In the computational theory of perceptions (CPT) (Zadeh 1999), words play the role of labels of perceptions and, more generally, perceptions are expressed as propositions in a natural language. Computing with words (CW) techniques are employed to translate propositions expressed in a natural language into what is called the generalized constraint language (GCL). In this language, the meaning of a proposition is expressed as a generalized constraint, X *isr* R, where X is the constrained variable, R is the constraining relation and *isr* is a variable copula in which r is a discrete variable whose value defines the way in which R constrains X (Zadeh 1997 and 1999). Among the basic types of constraints are : possibilistic, veristic, probabilistic, random set, Pawlak set, fuzzy graph and usuality (Zadeh 1999). These perceptions are mainly manipulated based on fuzzy concepts. For processing a fuzzy query, the PI is consisted of attributes associated with a keyword restricted by fuzzy term(s) in a fuzzy query. In this respect, the restricted keyword is named as a focal keyword, whereas attribute(s) associated with the focal keyword may be regarded as focal attribute(s). The PI can be mainly derived from the contents in the text body of a Web page or from the other sources of information with respect to a Web page. For example, the PI may be consisted of distance, size, weight, color, etc. on a keyword in the text body of a Web page. Using the PI, search engines can process fuzzy concepts (terms). In the sequel, if we integrate the DI used in commercial Web search engines with the proposed PI, search engines can process fuzzy queries. For example, consider a fuzzy query that finds '***popular*** national parks in the USA'. In this case, the fuzzy term '*popular*' is processed by using the PI, whereas keywords '*national parks*' and '*USA*' are processed by using the DI. We note that 'in' and 'the' in the above fuzzy query are examples of stop words ignored by search engines (see <www.google.com>).

Table 8. An example of Integrated Index (DI + PI)

Document Index (DI)	IPs	Perception Index (PI)				FPs (Results)
Keywords	URLs	Distance	Size	No. of visitors	...	Targeted URLs

(IPs : Intermediate Pointers; FPs : Final Pointers; URLs : Uniform Resource Locators)

It should be noted that fuzzy term(s) may be regarded as a constraint on a fuzzy query. For example, consider a fuzzy query that finds '***popular*** national parks in the USA'. In this case, the fuzzy term '***popular***' play the role of a constraint on the fuzzy query. In other words, using fuzzy term(s), Internet users can narrow thousands of hits to the few that users really want. In this respect, the PI provides help-

ful hints for targeting queries that users really want, and an invaluable personalized search.

The expressive power of conventional search engine query interfaces is relatively weak when restricted to keyword-based search (Kao et al. 2000). At present, commercial Web search engines based on the DI (i.e., keyword-based search engines) present limitations in modeling perceptual aspects of humans. In addition, they generally return a bunch of Web pages (or URLs) irrelevant to user's query. Although much Web search engines have been developed, they do not properly handle the fuzzy terms representing human's perception. In addition, they appear to have trouble with returning the targeted results. In order to tackle this problem, we integrate the DI used in commercial Web search engines with the proposed PI. In the proposed method, given a fuzzy query, search engine processes the fuzzy query based on the integrated index (DI + PI) as in **Figure 8.**

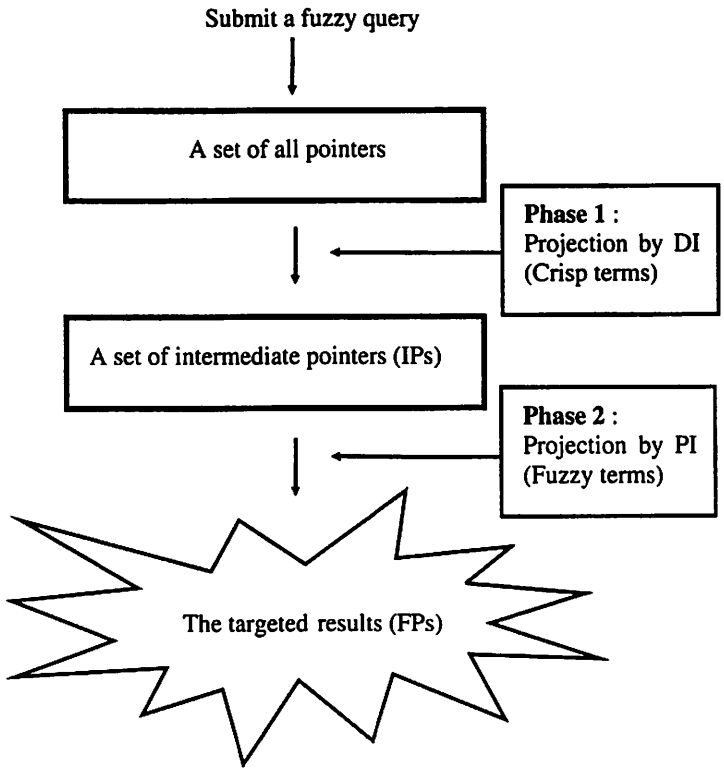


Figure 8. A search mechanism based on the integrated index (DI + PI)

In **Figure 8**, if we submit a query with only crisp terms (keyword-based query), this search engine uses only the phase 1. By applying the DI, the phase 1 performs an elimination-based approach to eliminate the URLs which are impossible to be the answers of the query. In this case, this search engine will return the same results that the existing search engines do. On the other hand, if we submit a query with both crisp terms and fuzzy terms, this search engine uses both phase 1 and phase 2. In this case, by applying the PI, the URLs reflecting fuzzy terms are extracted. More specifically, the phase 2 evaluates the fuzzy terms in detail on the set of intermediate pointers (i.e., the candidate URLs), and then generates the final pointers (FPs) (i.e., targeted results) that user really wants. This search mechanism can be conceptually explained by SQL-like language as follows : *SELECT * FROM [a set of intermediate pointers that satisfies focal keyword(s) in the DI] [WHERE the value(s) of focal attribute(s) in the PI are satisfied by the user].* We note that commercial Web search engines tend to ignore the importance of *[WHERE]* part. In this approach, the PI may be regarded as a constraint on the DI.

5.3 Fuzzy Query based on the Integrated Index (DI + PI)

We assume that a fuzzy term in a fuzzy query is marked with an asterisk. For example, it is expressed as '***popular*** national parks in the USA'. If a query has fuzzy term(s) marked with asterisk(s), search engine displays a PI associated with a focal keyword restricted by fuzzy term(s). Then user can specify values with respect to the fuzzy terms.

5.3.1 Types of Fuzzy Query

Fuzzy query is largely divided into simple fuzzy query and compound fuzzy query.

(1) Simple fuzzy query

The simple fuzzy query does not include conjunction ('*and*') or disjunction ('*or*') connective(s) between fuzzy terms, or negation ('*not*').

Example 1. Consider a fuzzy query that finds '**popular* national parks in the USA'. In this case, the DI, the PI, and stop words may be as follows : DI = {national parks, USA, ...}, PI = {No. of visitors, ...}, stop words = {in, the}. We note that a focal keyword 'national parks' in the DI is restricted by a fuzzy term 'popular'. In this case, the fuzzy term '*popular*' may be manipulated by the number of visitors (i.e., a focal attribute in the PI) per year, and represented as in **Figure 9**.

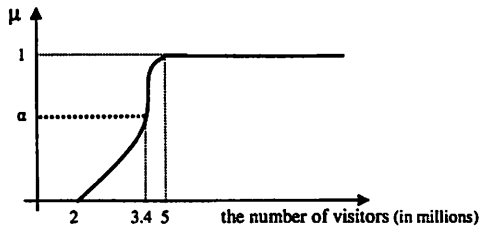


Figure 9. A membership function of '*popular*'

Example 2. Consider a fuzzy query that finds 'national parks ***moderate*** distance from San Francisco'. In this case, the DI, the PI and stop words may be as follows : DI = {national parks, San Francisco, ...}, PI = {distance, ...}, stop words = {from}. We note that a focal keyword 'San Francisco' in the DI is restricted by a fuzzy term 'moderate'. In this case, the fuzzy term '*moderate*' may be manipulated by the degree of distance (i.e., a focal attribute in the PI) from San Francisco, and represented as in **Figure 10.**

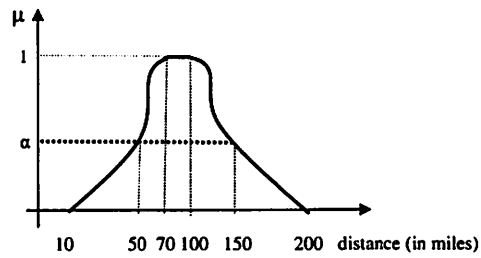


Figure 10. A membership function of '*moderate*'

(2) Compound fuzzy query

The compound fuzzy query includes conjunction ('*and*') or disjunction ('*or*') connective(s) between fuzzy terms, or negation ('*not*').

• ***Conjunction ('and')***

Example 3. Consider a fuzzy query that finds 'national parks that ***popular*** ***and*** ***moderate*** distance from San Francisco'. In this case, the DI, the PI, logical operator, and stop words may be as follows : DI = {national parks, San Francisco, ...}, PI = {No. of visitors, distance, ...}, logical operator = {and}, stop words =

{that, from}. We note that a focal keyword 'San Francisco' in the DI is restricted by fuzzy terms 'popular' and 'moderate'. In this case, the fuzzy terms '*popular*' and '*moderate*' may be manipulated as in **Figures 9** and **10**, respectively.

• ***Disjunction ('or')***

Example 4. Consider a fuzzy query that finds 'national parks that ***popular** or ***moderate** distance from San Francisco'. In this case, the DI, the PI, logical operator, and stop words may be as follows : DI = {national parks, San Francisco, ...}, PI = {No. of visitors, distance, ...}, logical operator = {or}, stop words = {that, from}. We note that a focal keyword 'San Francisco' in the DI is restricted by fuzzy terms 'popular' and 'moderate'. In this case, the fuzzy terms '*popular*' and '*moderate*' may be manipulated as in **Figures 9** and **10**, respectively.

• ***Negation ('not')***

Example 5. Consider a fuzzy query that finds '*not* ***popular*** national parks in the USA'. In this case, the DI, the PI, logical operator and stop words may be as follows : DI = {national parks, USA, ...}, PI = {No. of visitors, ...}, logical operator = {not}, stop words = {in, the}. This fuzzy query is similar to Example 1 but the fuzzy term 'popular' is negated. According to **Figures 9**, the negated fuzzy term '*not popular*' may be represented as in **Figure 11**.

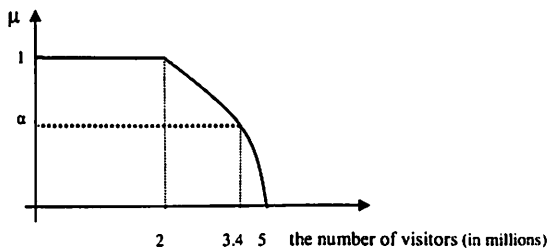


Figure 11. A membership function of '*not popular*'

5.3.2 Query Processing based on the Integrated Index (DI + PI)

Now, we present how this search engine processes fuzzy queries. Let the set of national parks in the USA be $A = \{A_1, A_2, \ldots, A_{99}, A_{100}\}$ and each A_i, (i = 1,2, ..., 100) has its own PT(page title) or URL.

Example 6. Consider a crisp query that finds 'national parks in the USA' (Q_1). In this case, the PI is not used. So, this search engine uses only the phase 1 in **Figure 8**. Thus, the integrated index is made as in **Table 9**.

Table 9. A snapshot of Integrated Index (DI + PI) after processing Q_1

Document Index (DI)	IPs	Perception Index (PI)			FPs (Results)
National parks, USA	A_1	Distance	No. of visitors	...	A_1
	A_2	Distance	No. of visitors	...	A_2
	...	...	...	...	...
	A_{99}	Distance	No. of visitors	...	A_{99}
	A_{100} + Irrelevant URLs	Distance	No. of visitors	...	A_{100} + Irrelevant URLs

In the crisp query case, this search engine returns the same results that the existing search engines do. We note that IPs and FPs are equal.

Example 7. Consider a fuzzy query that finds '***popular*** national parks in the USA' (Q_2). In this case, the DI and the PI are used. So, this search engine uses both the phase 1 and the phase 2 in **Figure 8**. We assume that '***popular*** national parks in the USA' are A_p, $A_p \in \{A_1, A_2, ..., A_{99}, A_{100}\}$, by using α-cut in **Figure 9**. Thus, the integrated index is made as in **Table 10**.

Table 10. A snapshot of Integrated Index (DI + PI) after processing Q_2

Document Index (DI)	IPs	Perception Index (PI)			FPs (Results)
National parks, USA	$\{A_1,...,A_{100}\}$ + Irrelevant URLs	Dis-tance	No. of visitors	...	URLs w.r.t $\{A_p\}$

(Focal keyword : National parks; Focal attribute : No. of visitors)

Example 8. Consider a fuzzy query that finds 'national parks ***moderate*** distance from San Francisco' (Q_3). In this case, the DI and the PI are used. So, this search engine uses both the phase 1 and the phase 2 in **Figure 8**. We assume that 'national parks ***moderate*** distance from San Francisco' are A_m, $A_m \in \{A_1, A_2, \ldots, A_{99}, A_{100}\}$, by using α-cut in **Figure 10**. Thus, the integrated index is made as in **Table 11**.

Table 11. A snapshot of Integrated Index (DI + PI) after processing Q_3

Document Index (DI)	IPs	Perception Index (PI)			FPs (Results)
National parks, **San Francisco**	$\{A_1,\ldots, A_{100}\}$ + Irrelevant URLs	**Distance**	No. of visitors	...	URLs w.r.t $\{A_m\}$

(Focal keyword : San Francisco; Focal attribute : Distance)

Example 9. Consider a fuzzy query that finds 'national parks that ***popular*** *and* ***moderate*** distance from San Francisco' (Q_4). In this case, the DI and the PI are used. So, this search engine uses both the phase 1 and the phase 2 in **Figure 8**. Then the query results with respect to Q_4 become $\{A_p\} \cap \{A_m\}$. For instance, let A_p be a set $\{A_1, A_2, A_3\}$ and A_m be a set $\{A_1, A_4, A_5\}$, then $\{A_p\} \cap \{A_m\} = \{A_1\}$. Thus, the integrated index is made as in **Table 12**.

Table 12. A snapshot of Integrated Index (DI + PI) after processing Q_4

Document Index (DI)	IPs	Perception Index (PI)			FPs (Results)
National parks, **San Francisco**	$\{A_1,\ldots, A_{100}\}$ + Irrelevant URLs	**Distance**	**No. of visitors**	...	URLs w.r.t $\{A_p\} \cap \{A_m\}$

(Focal keyword : San Francisco; Focal attributes : Distance and no. of visitors)

Example 10. Consider a fuzzy query that finds 'national parks that ***popular** or ***moderate** distance from San Francisco' (Q_5). In this case, the DI and the PI are used. So, this search engine uses both the phase 1 and the phase 2 in **Figure 8**. Then the query results with respect to Q_5 become $\{A_p\} \cup \{A_m\}$. For instance, let A_p be a set $\{A_1, A_2, A_3\}$ and A_m be a set $\{A_1, A_4, A_5\}$, then $\{A_p\} \cup \{A_m\} = \{A_1, A_2, A_3, A_4, A_5\}$. Thus, the integrated index is made as in **Table 13**.

Table 13. A snapshot of Integrated Index (DI + PI) after processing Q_5

Document Index (DI)	IPs	Perception Index (PI)			FPs (Results)
National parks, **San Francisco**	$\{A_1, \ldots, A_{100}\}$ + Irrelevant URLs	**Distance**	**No. of visitors**	...	URLs w.r.t $\{A_p\} \cup \{A_m\}$

(Focal keyword : San Francisco; Focal attributes : Distance and no. of visitors)

Example 11. Consider a fuzzy query that finds '*not* ***popular** national parks in the USA' (Q_6). In this case, the DI and the PI are used. So, this search engine uses both the phase 1 and the phase 2 in **Figure 8**. Then the query results with respect to Q_6 become $\{\sim A_p\}$. For instance, let A_p be a set $\{A_1, A_2, A_3\}$, then $\{\sim A_p\} = \{A_4, A_5, \ldots, A_{99}, A_{100}\}$ if the universal set $A = \{A_1, A_2, \ldots, A_{99}, A_{100}\}$. Thus, the integrated index is made as in **Table 14**.

Table 14. A snapshot of Integrated Index (DI + PI) after processing Q_6

Document Index (DI)	IPs	Perception Index (PI)			FPs (Results)
National parks, USA	$\{A_1, \ldots, A_{100}\}$ + Irrelevant URLs	Distance	**No. of visitors**	...	URLs w.r.t $\{\sim A_p\}$

(Focal keyword : National parks; Focal attribute : No. of visitors)

5.3.3 User Interface based on the Integrated Index (DI + PI)

Williams (1984) developed a user interface for information retrieval systems to aid users in formulating a query. The system, *RABBIT III*, supports interactive refinement of queries by allowing users to critique retrieved results with labels such as 'require' and 'prohibit'. Williams claims that this system is particularly helpful to naïve users with only a vague idea of what they want and therefore need to be guided in the formulation/reformulation of their queries or who have limited knowledge of a given database or who must deal with a multitude of databases. This process allows users to refine their queries. In a similar sense, we can refine user's query by means of the phase 2 for processing fuzzy term(s) in **Figure 8**. Thus, search engine will return the targeted results that users really want. An important problem relating to personalization concerns understanding how a machine can help an individual user via suggesting recommendations (Bekin 2000). In our approach, the PI can help the user to specify clearly what he/she really wants. More specifically, the user in the system is asked to specify fuzzy term(s) in a query. In this respect, the PI may be regarded as a recommendation for handling fuzzy term(s) in a query. As a result, search engine returns 'the targeted results'. Now, we describe user interface for phase 1 and phase 2 in **Figure 8**.

(1) User interface for phase 1

Initially, user interface for phase 1 lets user specify his/her queries with only crisp terms (keywords), or both crisp terms and fuzzy terms. If user submits a query with only crisp terms, only user interface for phase 1 is used, and search results are returned based on only the DI. On the other hand, if user submits a query with both crisp terms and fuzzy terms, user interface for phase 2 is also displayed to process the fuzzy terms.

(2) User interface for phase 2

For the fuzzy query on the Internet, the 'easy of use' is important because Internet users are broad spectrum in terms of cultural differences, level of intelligence, etc. In this respect, user interface for phase 2 should provide Internet users with an easy user interface for specifying these fuzzy terms such as '*popular*', '*moderate*', '*big*', etc. In addition, we need to reflect cultural differences. For instance, different people generally use different scales (i.e., feet, miles, meter, etc). Internet users have their own membership functions with respect to fuzzy terms in a fuzzy query, by means of human's perception capability. Consequently, they can give values with respect to fuzzy terms in the user interface for phase 2.

User interface for phase 2 displays a PI associated with a focal keyword. For example, given a fuzzy query that finds '***popular*** national parks in the USA', a PI

associated with a focal keyword 'national parks' is displayed as shown in **Table 10.** It should be noted that different people may use different conceptual comprehension (fuzzy terms, membership functions, α-cut), with respect to the same situation. It is the user's task in this user interface to examine the suggested attributes in the PI, and to specify the values of the focal attributes reflecting user's query requirements. Using the PI, search results can be restricted within narrow limit. We call it '*target search by fuzzy terms*'. In other words, search engine will return the targeted results that users really want.

Fuzzy terms are specified in user interface for phase 2. For example, they can be expressed as point value, interval value, multiple values, etc.

- ***Point value***

Example 12. In Example 1, given a α-cut, the fuzzy term '*popular*' may be specified by using a focal attribute 'no. of visitors'. More specifically, it is expressed as a point 3.4 (i.e., 'no. of visitors' $\geq$ 3.4 millions).

- ***Interval value***

Example 13. In Example 2, given a α-cut, the fuzzy term '*moderate*' may be specified by using a focal attribute 'distance'. More specifically, it is expressed as an interval (i.e., distance = [50, 150] in miles).

- ***Multiple values***

A veristic variable (Zadeh 1997 and 1999) which can be assigned two or more values in its universe simultaneously will be specified as multiple values.

Example 14. Let U be the universe of natural languages and let X denote the fluency of an individual in English, French and Italian. Then, X *isv* (1.0 English + 0.8 French + 0.6 Italian) means that the degrees of fluency of X in English, French and Italian are 1.0, 0.8 and 0.6, respectively (Zadeh 1997 and 1999).

5.4 Some Features of the Proposed Method

Remark 1. The higher the α in α-cut ($0 \leq \alpha \leq 1$), the smaller the number of the targeted results. This property provides continual incremental result from 'the highest constraint (i.e., $\alpha = 1$)' to 'the lowest constraint (i.e., $\alpha = 0$)'. Consequently, we can achieve 'interactive user control of the query processing' by adjusting the value of α.

Remark 2. If $\alpha = 0$, search results coincide with the results by applying only the DI (i.e., the existing keyword-based search). In this case, the results of phase 1 in **Figure 8** become search results.

Remark 3. Even though the same integrated index (DI+PI) is given, different search results are returned by adjusting the value of α or by using different focal attributes in the PI. In the case of 'using different focal attributes in the PI', for example, consider a fuzzy query that finds '*attractive* car', where '*attractive*' means 'comfortable and fast'. In this case, for the fuzzy term '*attractive*', people may use different focal attributes (i.e., size, speed, etc.) in the PI. In addition, different people may use different conceptual comprehension (fuzzy terms, membership functions, α-cut), with respect to the same situation. Thus, search engine will return the personalized search results that users really want. In the meantime, clustering (i.e., grouping similar documents together to expedite information retrieval) is adaptively determined depending on the value of α or the selected focal attributes in the PI.

Remark 4. Using the PI, Internet users can narrow thousands of hits to the few that users really want.

Remark 5. Using the PI, therefore, we can tackle in part the major problems in commercial Web search engines (i.e., 'large answer set', 'low precision', 'ineffective for general-concept queries').

5.5 Performance Analysis

For comparing with commercial keyword-based search engines, the ratio *[the number of FPs / the number of IPs]* can be used as a measure of performance evaluation on the proposed method. We note that the number of IPs is the result of phase 1 and the number of FPs is the result of phase 2 in **Figure 8**. The smaller the ratio, the better the filtering effect of the proposed method. More specifically, **Table 15** illustrates the problem 'quality of retrieved information' in the commercial Web search engines by showing the results obtained from querying two popular search engines with 6 sample queries ($Q_1 \sim Q_6$ in Subsec. 5.3.2).

Table 15. Example queries and results

Queries	Search engines	No. of hits
Q_1 (Crisp Query)	Yahoo !	returns about 112,000
	Google	returns about 240,000
	The proposed method	returns the same results that the existing search engines do
Q_2 (Fuzzy Query)	Yahoo !	returns about 34,200
	Google	returns about 73,100
	The proposed method	returns URLs w.r.t $\{A_p\}$ (see **Table 10**)
Q_3 (Fuzzy Query)	Yahoo !	returns about 1,380
	Google	returns about 3,960
	The proposed method	returns URLs w.r.t $\{A_m\}$ (see **Table 11**)
Q_4 (Fuzzy Query)	Yahoo !	returns about 1,330
	Google	returns about 2,050
	The proposed method	returns URLs w.r.t $\{A_p\} \cap \{A_m\}$ (see **Table 12**)
Q_5 (Fuzzy Query)	Yahoo !	returns about 1,000
	Google	returns about 2,050
	The proposed method	returns URLs w.r.t $\{A_p\} \cup \{A_m\}$ (see **Table 13**)
Q_6 (Fuzzy Query)	Yahoo !	returns about 29,100
	Google	returns about 62,200
	The proposed method	returns URLs w.r.t $\{\sim A_p\}$ (see **Table 14**)

5.6 Additional Considerations

The work of Lidsky and Kwon (1997) is an opinionated but informative resource on search engines. It describes 36 different search engines and rates them on specific details of their search capabilities. For instance, in one study, searches are divided into five categories : (1) simple searches; (2) custom searches; (3) directory searches; (4) current news searches; and (5) Web content. The five categories of search are evaluated in terms of power and easy of use. Variations in ratings sometimes differ substantially for a given search engine. In the meantime, they chose the respective best search engine according to five categories : (1) search indexes and directories; (2) people finders; (3) business finders; (4) usenet search; and (5) metasearch. The data indicate that as the number of people using the Internet and Web has grown, user types have diversified and search engine providers have begun to target more specific types of users and queries with specialized and tailored search tools. In this respect, for the fuzzy query processing, topic-specific (or domain-specific) requirement is necessary because of the following reasons : (1) commonsense knowledge - the present state of AI is not up to formulating a full commonsense database, but full commonsense knowledge is not necessary (McCarthy 2000). In this respect, for the fuzzy query processing, if we design a search engine based on 'domain-specific' concept, the degree of freedom on fuzzy terms will be highly reduced. In other words, 'domain-specific' concept provides the higher possibility for a well-defined (restricted) condition. For example, given a travel-domain database, consider a fuzzy query that finds '***popular** national parks in the USA'. In this case, the fuzzy term 'popular' is used to restrict 'national parks', not 'music', 'car', etc.; (2) indexing overhead - human indexing (for example, Yahoo !, LookSmart, etc.) is currently the most accurate because experts on popular subjects organize and compile the directories and indexes in a way which facilitates the search process. However, the enormous number of existing Web pages and their rapid increase and frequent updating make the indexing a difficult one or an overhead. If we design a search engine based on 'domain-specific' concept, the indexing overhead on the PI will be highly reduced; (3) storage requirement - comparing with traditional Web search engines, recommending the PI requires the system to maintain more data. If we design a search engine based on 'domain-specific' concept, the storage requirement on the PI will be highly reduced; (4) uneven concentration - if we design a search engine based on 'domain-specific' concept, 'uneven concentration of information packets on the Internet' problem, as described in Section 1, will be highly reduced.

5.7 Remarks

Although the commercial Web search engines such as Yahoo !, Google, Lycos, etc. help Internet users get to good information, they do not properly handle fuzzy query. For example, consider a fuzzy query that finds '***popular*** national parks in the USA'. In this case, '***popular***', 'national parks' and 'USA' are generally proc-

essed as keywords in the commercial Web search engines. It should be noted that fuzzy term '***popular***' is a constraint on a focal keyword 'national parks' rather than an independent keyword. In other words, the fuzzy term '***popular***' plays the role of a constraint on the fuzzy query. However, commercial Web search engines tend to ignore the importance of fuzzy terms in a query processing. As a result, search engines return a bunch of page titles (or URLs) irrelevant to user's query. For example, in the case of a fuzzy query that finds '***popular*** national parks in the USA', Yahoo ! returns about 34,200 page titles (or URLs) and Google returns about 73,100 page titles (or URLs). Intuitively, we find that there are so many page titles (or URLs) irrelevant to user's query.

In this section, we present the search mechanism based on the integrated index (DI + PI) and fuzzy query based on the integrated index (DI + PI). Moreover, we describe some features of the proposed method and suggest some considerations for implementing the proposed method.

6 Ranking Algorithm based on Perception Index

In Section 5, we have introduced the Perception Index (PI) that contains attributes associated with a focal keyword restricted by fuzzy term(s) in a fuzzy query.

Ranking algorithms play an important role in Web search engines. Although the existing ranking methods for Web search engines provide users with their own ranking algorithms based on popularity, bookmark, etc., they often tend to return unranked random samples in response to user's query. In order to tackle this problem, we introduce a new ranking algorithm based on the Perception Index (PI). Using the values of focal attributes in the PI, user's search intentions can be explicitly reflected. Consequently, the proposed ranking algorithm ensures consistently high-quality returns in terms of user's search intentions.

6.1 Overview

Increased capabilities of computer hardware and software have created a vast body of machine-readable resources. Typically there is no lack of available information; more often, users, seeking needles in haystacks, are overwhelmed by the quantity of irrelevant information. Often this is caused by a poor query (too vague or too generic; for example, try searching for "computer science") (Ali and McRoy 2000) . Without the context of the query and the relations of the information, a search engine is doomed to return random samples of the Internet. With no ability to control or organize the sprawl on the Internet, how will we ever be able to find the intelligence in all the data, information and knowledge that we presume

to be there ? Perhaps the Internet is more like TV. Is it mostly a collection of garbage, gleaned at the lowest common denominator, serving merely to provide eyeballs to advertisers; or is it a free exchange of information that's just too cheap to meter ? (Hoebel and Welty 1999). Despite numerous refinements, most Web search engines still return too many results and random samples of the Internet. In other words, they often give users a bunch of garbage. In this respect, we need a new tool to handle both the removal of spurious results and the random samples of the Internet. In section 5, we have mainly discussed the problem on 'the removal of spurious results'. The PI provides a deductive capability to query language on the Internet. In other words, the useful URLs (targeted URLs) are separated from the useless by the PI. In this section, we will focus on the ranking within the targeted URLs. The Compaq study found that most searchers (68%) look only at the first page of results. This means that ranking algorithm plays an important role in Web search engines. Although the existing ranking methods for Web search engines provide users with their own ranking algorithms based on popularity, bookmark, etc., they often tend to return unranked random samples in response to user's query. In order to tackle this problem, we introduce a new ranking algorithm based on the Perception Index (PI). Using the values of focal attributes in the PI, user's search intentions can be explicitly reflected. For example, consider a fuzzy query that finds '*attractive* car', where '*attractive*' means 'comfortable and fast'. In this case, for the fuzzy term '*attractive*', people may use different focal attributes (i.e., size, speed, etc.) in the PI. In this respect, it provides a user with the personalized ranking based on user's search intentions.

In Section 6.2, we briefly summarize the existing ranking methods. In Section 6.3, we introduce a new ranking algorithm based on the PI, and compare the proposed ranking algorithm with the existing Web ranking methods. We discuss the proposed ranking method in Section 6.4.

6.2 Summary of the existing ranking methods

In conventional information retrieval (IR), a variety of techniques have been developed for ranking retrieved documents for a given query. A textual database can be represented by a word-by-document matrix whose entries represent the frequency of occurrence of a word in a document. Thus, documents can be thought of as vectors in a multidimensional space, the dimensions of which are the words used to represent the texts. In a standard 'keyword-matching' vector system (Salton and McGill 1983), the similarity between two documents is computed as the inner product or cosine of the corresponding two columns of the word-by-document matrix. Queries can also be represented as vectors of words and thus compared against all document columns with the best matches being returned. An important assumption in this vector space model is that the words (i.e., dimensions of the space) are orthogonal or independent. While it has been a reasonable first approximation, the assumption that words are pairwise independent is not realistic.

Recently, several statistical and AI techniques have been used to better capture term association and domain semantics. One such method is latent semantic indexing (LSI) (Berry et al. 1995, Deerwester et al. 1990). LSI is an extension of the standard vector retrieval method designed to help overcome some of the retrieval problems described previously. In LSI the associations among terms and documents are calculated and exploited in retrieval. The assumption is that there is some underlying or 'latent' structure in the pattern of word usage across documents and that statistical techniques can be used to estimate this latent structure. A description of terms, documents, and user queries based on the underlying latent semantic structure is used for representing and retrieving information. The particular LSI analysis described by Deerwester et al. (1990) uses singular value decomposition (SVD), a technique closely related to eigenvector decomposition and factor analysis. SVD takes a large word-by-document matrix and decomposes it into a set of *k*, typically 100 to 300, orthogonal factors from which the original matrix can be approximated by linear combination. Instead of representing documents and queries directly as vectors of independent words, LSI represents them as continuous values on each of the *k* orthogonal indexing dimensions derived from the SVD analysis. One advantage of this approach is that queries can retrieve documents even if they have no words in common. The LSI technique captures deeper associative structure than simple term-to-term correlations and clusters and is completely automatic. We can interpret the analysis performed by SVD geometrically. The result of the SVD is a *k*-dimensional vector space containing a vector for each term and each document. The location of term vectors reflects the correlations in their usage across documents. Similarly, the location of document vectors reflects correlations in term usage. In this space the cosine or dot product between vectors corresponds to their estimated similarity. Retrieval proceeds by using the terms in a query to identify a vector in the space, and all documents are then ranked by their similarity to the query vector. The LSI method has been applied to several standard IR collections with favorable results.

In the Web search engines, however, detailed information regarding ranking algorithms used by major search engines is not publicly available. A simple means to measure the quality of a Web page, proposed by Carriere and Kazman (1997), is to count the number of pages with pointers to the page. Google is a representative Web search engine that uses link information. Its rankings are based, in part, on the number of other pages with pointers to the page. In November 1999, Northern Light introduced a new ranking system, which is also based, in part, on link data (see <http://www.searchenginewatch.com/sereport/ 99/11briefs.html>). In other words, Google and Northern Light rank search results, in part, by popularity. In the meantime, HotLinks ranks search results based on the bookmarks of its registered users. Yahoo's Inktomi-served results aren't ranked by popularity (see <http://websearch.about.com/internet/webserch/library /weekly/aa052199.htm>).

In theory, more popular links indicate more relevant content, but if a user differs from the crowd, simply popularity-based ranking approaches dive deeply into other possibilities on the Web. Consequently, they often give users a bunch of

garbage. In addition, they often tend to return unranked random samples in response to user's query. In order to tackle these problems, we introduce a new ranking algorithm based on the Perception Index (PI).

6.3 A new ranking algorithm based on the Perception Index (PI)

As described in Section 6.2, the existing ranking methods can be largely categorized into the following two classes : keyword-based approach and hyperlink-based approach. On the other hand, in Section 5, we have shown that fuzzy terms play the role of a constraint on the fuzzy query and can be expressed by using the focal attributes in the PI. In this Section, we propose a new ranking algorithm based on the PI. It may be regarded as a fuzzy term-based approach.

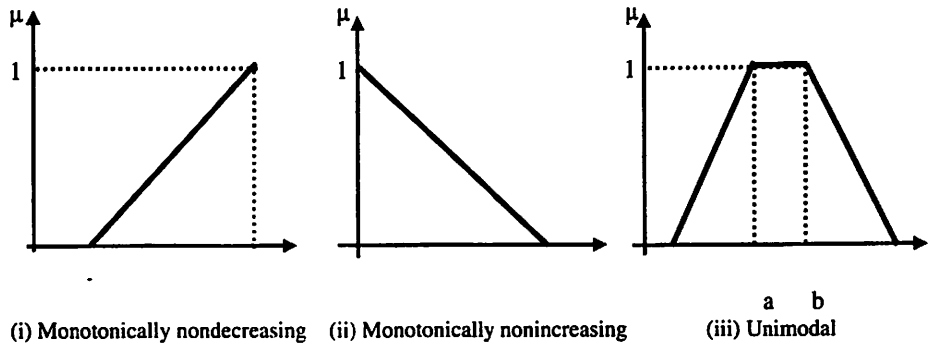
Although Web search engines generally return large amounts of web pages (or URLs) for a given query, only a small fraction of the returns will actually be relevant to any particular person. Thus, there is the problem of determining what information is of interest to any particular person, while minimizing the amount of search through irrelevant information. In Section 5, we have mainly discussed the problem on 'the removal of spurious results' irrelevant to user's search intentions. The PI provides a deductive capability to query language on the Internet. In other words, the useful URLs (targeted URLs) are separated from the useless by the PI. In this Section, we will focus on the ranking within the targeted URLs. The Compaq study found that most searchers (68%) look only at the first page of results. This means that ranking algorithm plays an important role in Web search engines. Although the existing ranking methods for Web search engines provide users with their own ranking algorithms based on popularity, bookmark, etc., they often tend to return unranked random samples in response to user's query. In order to tackle this problem, we introduce a new ranking algorithm based on the PI. It provides a user with the personalized ranking based on user's search intentions.

Zadeh suggested we can represent linguistic quantifiers as fuzzy subsets of the unit interval (Zadeh 1997). In this representation the membership grade of any proportion $r \in [0, 1]$, $Q(r)$, is a measure of the compatibility of the proportion r with the linguistic quantifier we are representing by the fuzzy subset Q. For example, if Q is the quantifier 'most' then Q(0.9) represents the degree to which 0.9 satisfies the concept 'most'. Yager identified three classes of linguistic quantifiers that cover most of these used in natural language (Yager 1991 and 1996).

(i) A quantifier Q is said to be monotonically nondecreasing if $r_1 > r_2$ then $Q(r_1) \geq Q(r_2)$.
(ii) A quantifier Q is said to be monotonically nonincreasing if $r_1 > r_2$ then $Q(r_1) \leq Q(r_2)$.

(iii) A quantifier Q is said to be unimodal if there exists two values $a \leq b$ both contained in the unit interval such that for $r < a$, Q is monotonically nondecreasing, for $r > b$, Q is monotonically nonincreasing, and for $r \in [a, b]$, $Q(r) = 1$.

Figure 12 shows prototypical examples of these quantifiers.



(i) Monotonically nondecreasing (ii) Monotonically nonincreasing (iii) Unimodal

Figure 12. Three types of quantifiers

In a similar way, we can identify three classes of fuzzy terms that cover most of these used in natural language. For example, in Section 5.3.1, we have represented the fuzzy terms 'popular' (monotonically nondecreasing), 'moderate' (unimodal), 'not popular' (monotonically nonincreasing), (see **Figures 9-11**). In this respect, we design a new ranking algorithm based on the PI as follows :

Algorithm 1 : *Ranking for one focal attribute*

(i) Monotonically nondecreasing case

The larger the value of focal attribute in the PI, the higher the rank retrieved documents for a given fuzzy query.

(ii) Monotonically nonincreasing case

The larger the value of focal attribute in the PI, the lower the rank retrieved documents for a given fuzzy query.

(iii) Unimodal case

If an interval of focal attribute determined by α-cut is $[a_i, b_i]$, and let β denote the midpoint between a_i and b_i, then the degree of closeness (nearness) to β can be used as a ranking criterion. In other words, the closer the β, the higher the rank retrieved documents for a given fuzzy query.

Example 15. Consider a fuzzy query that finds '***popular*** national parks in the USA'. In this case, the fuzzy term 'popular' may be represented by a monotonically nondecreasing membership function, (see **Figure 9**)We assume that '***popular*** national parks in the USA' are A_p by using α-cut. Let the targeted results A_p be $\{ A_p^1, A_p^2, ..., A_p^r \}$ taking values of focal attribute (i.e., no. of visitors) such as $\text{Val}(A_p^1) \le \text{Val}(A_p^2) \le ... \le \text{Val}(A_p^r)$, then the targeted results A_p are ranked as the following order: $A_p^r, ..., A_p^2, A_p^1$.

Example 16. Consider a fuzzy query that finds 'national parks ***moderate*** distance from San Francisco'. In this case, the fuzzy term 'moderate' may be represented by a unimodal membership function, (see **Figure 10**). We assume that '***moderate*** national parks in the USA' are A_m by using α-cut. Let an interval of focal attribute (i.e., distance) determined by α-cut be $[a_i, b_i]$, and let β denote the midpoint between a_i and b_i, and let the targeted results A_m be $\{ A_m^1, A_m^2, ..., A_m^s \}$ taking values of the focal attribute such as $\text{Val}(A_m^1)$, $\text{Val}(A_m^2)$, ..., $\text{Val}(A_m^s)$. If the degree of closeness (nearness) to β is the order $\text{Val}(A_m^1)$, $\text{Val}(A_m^2)$, ..., $\text{Val}(A_m^s)$, then the targeted results A_m are ranked as the following order: $A_m^1, A_m^2, ..., A_m^s$.

Example 17. Consider a fuzzy query that finds 'not ***popular*** national parks in the USA'. In this case, the negated fuzzy term 'not popular' may be represented by a monotonically nonincreasing membership function, (see **Figure 11**). We assume that 'not ***popular*** national parks in the USA' are $\sim A_p$ by using α-cut. Let the targeted results $\sim A_p$ be $\{ A_p^1, A_p^2, ..., A_p^t \}$ taking values of focal attribute (i.e., no. of visitors) such as $\text{Val}(A_p^1) \le \text{Val}(A_p^2) \le ... \le \text{Val}(A_p^t)$, then the targeted results $\sim A_p$ are ranked as the following order: $A_p^1, A_p^2, ..., A_p^t$.

Algorithm 2 : ***Ranking for multiple focal attributes***

If we have multiple focal attributes (for instance, 'no. of visitors' and 'distance'), weighting the importance of focal attributes should be considered. For the weighted case, assume that $\theta_1, \theta_2, \ldots, \theta_n$ are ordinal weights. Then we refer to $\Theta = (\theta_1, \theta_2, \ldots, \theta_n)$ as a weighting, where θ_i is the weight of attribute i. Intuitively, the targeted results can be ranked according to the ordinal weights. For a respective focal attribute, the rank retrieved documents for a given fuzzy query can be determined based on the **Algorithm 1.**

Example 18. Consider a fuzzy query that finds 'national parks that ***popular and moderate*** distance from San Francisco'. In this case, the fuzzy terms 'popular' and 'moderate' may be represented by a monotonically nondecreasing membership function and a unimodal membership function, respectively. Using the results of Examples 15 and 16, if the weight of focal attribute 'no. of visitors' is more important than the weight of focal attribute 'distance', then the targeted results are ranked as the following order : $A_p^r, \ldots, A_p^2, A_p^1, A_m^1, A_m^2, \ldots, A_m^s$.

As described in Section 5.3.3, user interface for phase 2 displays a PI associated with a focal keyword. For example, consider a fuzzy query that finds '***popular*** national parks in the USA', a PI associated with a focal keyword 'national parks' is displayed. It is the user's task in this user interface to examine the suggested attributes in the PI, and to specify the values of the focal attributes reflecting user's search intentions. Consequently, search results can be restricted within narrow limit. We call it '*target search by fuzzy terms*'. In other words, search engine will return the targeted results that users really want. Now, if we apply Algorithm 1 and 2, the targeted results can be displayed from the highest rank to the lowest rank.

Although the existing ranking methods for Web search engines also provide users with their own ranking algorithms based on popularity, bookmark, etc., their approaches look like the behind-the-scenes processing. In the proposed approach, user's search intentions can be explicitly reflected by using the values of focal attributes in the PI. In this respect, we can explicitly describe how to rank the search results by means of the proposed approach. Consequently, the proposed approach provides a user with the personalized ranking based on user's search intentions.

7 Challenges and Road Ahead

During the August 2001, BISC program hosted a workshop toward better understanding of the issues related to the Internet (Fuzzy Logic and the Internet-FLINT2001, Toward the Enhancing the Power of the Internet). The main purpose of the Workshop was to draw the attention of the fuzzy logic community as well as the Internet community to the fundamental importance of specific Internet-related problems. This issue is critically significant about problems that center on search and deduction in large, unstructured knowledge bases. The Workshop provided a unique opportunity for the academic and corporate communities to address new challenges, share solutions, and discuss research directions for the future. Followings are the areas that were recognized as challenging problems and the new direction toward the next generation of the search engines and Internet. We summarize the challenges and the road ahead into four categories as follows:

I. Search Engine and Queries:

- Deductive Capabilities
- Customization and Specialization
- Metadata and Profiling
- Semantic Web
- Imprecise-Querying
- Automatic Parallelism via Database Technology
- Approximate Reasoning
- Ontology
- *Ambiguity Resolution through Clarification Dialog; Definition/Meaning & Specificity* User Friendly
- Multimedia
- Databases
- Interaction

II. Internet and the Academia:

- Ambiguity and Conceptual and Ontology
- Aggregation and Imprecision Query
- Meaning and structure Understanding
- Dynamic Knowledge
- Perception, Emotion, and Intelligent Behavior
- Content-Based
- Escape from Vector Space Deductive Capabilities
- Imprecise-Querying
- *Ambiguity Resolution through Clarification Dialog*
- *Precisiated Natural Languages (PNL)*

III. Internet and the Industry:

- XML=>Semantic Web
- Workflow
- Mobile E-Commerce
- CRM
- Resource Allocation
- Intent
- Ambiguity Resolution
- Interaction
- Reliability
- Monitoring
- Personalization and Navigation
- Decision Support
- Document Soul
- Approximate Reasoning
- Imprecise Query
- Contextual Categorization

IV. Fuzzy Logic and Internet; Fundamental Research:

- Computing with Words (CW)
- Computational Theory of Perception (CTP)
- Precisiated Natural Languages (PNL)

The potential Area and applications of Fuzzy Logic for the Internet include:

I. Potential Areas:

- Search Engines
- Retrieving Information
- Database Querying
- Ontology
- Content Management
- Recognition Technology
- Data Mining
- Summarization
- Information Aggregation and Fusion
- E-Commerce
- Intelligent Agents
- Customization and Personalization

I. Potential Applications:

- Search Engines and Web Crawlers
- Agent Technology (i.e., Web-Based Collaborative and Distributed Agents)
- Adaptive and Evolutionary techniques for dynamic environment (i.e. Evolutionary search engine and text retrieval, Dynamic learning and adaptation of the Web Databases, etc)
- Fuzzy Queries in Multimedia Database Systems
- Query Based on User Profile
- Information Retrievals
- Summary of Documents
- Information Fusion Such as Medical Records, Research Papers, News, etc
- Files and Folder Organizer
- Data Management for Mobile Applications and eBusiness Mobile Solutions over the Web
- Matching People, Interests, Products, etc
- Association Rule Mining for Terms-Documents and Text Mining
- E-mail Notification
- Web-Based Calendar Manager
- Web-Based Telephony
- Web-Based Call Centre
- Workgroup Messages
- E-Mail and Web-Mail
- Web-Based Personal Info
- Internet related issues such as Information overload and load balancing, Wireless Internet-coding and D-coding (Encryption), Security such as Web security and Wireless/Embedded Web Security, Web-based Fraud detection and prediction, Recognition, issues related to E-commerce and E-bussiness, etc.

8 Conclusion

Intelligent search engines with growing complexity and technological challenges are currently being developed. This requires new technology in terms of understanding, development, engineering design and visualization. While the technological expertise of each component becomes increasingly complex, there is a need for better integration of each component into a global model adequately capturing the imprecision and deduction capabilities. In addition, intelligent models

can mine the Internet to conceptually match and rank homepages based on predefined linguistic formulations and rules defined by experts or based on a set of known homepages. The FCM model can be used as a framework for intelligent information and knowledge retrieval through conceptual matching of both text and images (here defined as "Concept"). The FCM can also be used for constructing fuzzy ontology or terms related to th context of the query and search to resolve the ambiguity. This model can be used to calculate conceptually the degree of match to the object or query.

The expressive power of conventional search engine query interfaces is relatively weak when restricted to keyword-based search (i.e., Document Index (DI)-based search). At present, the keyword-based search engines present limitations in modeling perceptual aspects of humans. In addition, they appear to have trouble with returning the targeted results. In other words, they generally return a bunch of Web pages (or URLs) irrelevant to user's query. In this respect, we need a new tool to handle both the fuzzy query and the removal of spurious results. In order to tackle these problems, we introduce the Perception Index (PI) that contains attributes associated with a focal keyword restricted by fuzzy term(s) in a fuzzy query. If we integrate the Document Index (DI) used in commercial Web search engines with the proposed PI, we can handle both crisp terms (keyword-based) and fuzzy terms (perception-based). In this respect, the proposed approach is softer than the keyword-based approach (i.e., commercial Web search engines). It is a further step toward a real human-friendly, natural language-based interface for Internet. It should greatly help the user relatively easily retrieve relevant information. In other words, the proposed method assists the user to reflect his/her perception in the process of query. As a consequence, Internet users can narrow thousands of hits to the few that users really want. In this respect, the PI provides a new tool for targeting queries that users really want, and an invaluable personalized search. The use of PI provides helpful hints for solving the problems of 'large answer set', 'low precision', 'ineffective for general-concept queries' suffered by most search engines.

Although the existing ranking methods for Web search engines provide users with their own ranking algorithms based on popularity, bookmark, etc., they often tend to return unranked random samples in response to user's query. In theory, more popular links indicate more relevant content, but you will have to determine that for yourself. If you differ from the crowd, simply popularity-based ranking approaches dive deeply into other possibilities on the Web. Consequently, they often give users a bunch of garbage. In order to tackle this problem, we introduce a new ranking algorithm based on the Perception Index (PI). Using the values of focal attributes in the PI, user's search intentions can be explicitly reflected. For example, consider a fuzzy query that finds '*attractive* car', where '*attractive*' means 'comfortable and fast'. In this case, for the fuzzy term '*attractive*', people may use different focal attributes (i.e., size, speed, etc.) in the PI. In this respect, it provides a user with the personalized ranking based on user's search intentions. Conse-

quently, the proposed ranking algorithm ensures consistently high-quality returns in terms of user's search intentions.

Acknowledgement

Funding for this research was provided by the British Telecommunication (BT) and the BISC Program of UC Berkeley. This work was also supported by postdoctoral fellowships program from Korea Science & Engineering Foundation (KOSEF). The Authors thanks Sun-Gyung Jung, plan & control manager/education center of Oracle Korea.

References

S. S. Ali and S. McRoy (2000) Information retrieval, Intelligence (ACM) 11(4) : 17-19.

J. Baldwin, Future directions for fuzzy theory with applications to intelligent agents, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 200.

J. F. Baldwin and S. K. Morton, conceptual Graphs and Fuzzy Qualifiers in Natural Languages Interfaces, 1985, University of Bristol.

M. J. M. Batista et al., User Profiles and Fuzzy Logic in Web Retrieval, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

N.J. Belkin (2000) Helping people find what they don't know, Communications of the ACM 43(8) : 58-61.

H. Beremji, Fuzzy Reinforcement Learning and the Internet with Applications in Power Management or wireless Networks, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

M. W. Berry, S. T. Dumais and G. W. O'Brien (1995) Using linear algebra for intelligence information retrieval, SIAM Rev. 37(4) : 573-595.

T.H. Cao, Fuzzy Conceptual Graphs for the Semantic Web, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

J. Carriere and R. Kazman (1997) WebQuery : searching and visualizing the Web through connectivity, Proceedings of the sixth international conference on the World Wide Web.

D. Y. Choi, Integration of Document Index with Perception Index and Its Application to Fuzzy Query on the Internet, in M. Nikravesh and B. Azvine, FLINT 2001, New Di-

rections in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman (1990) Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41(6) : 391-407.

N. Guarino, C. Masalo, G. Vetere, "OntoSeek : content-based access to the Web", IEEE Intelligent Systems, Vol.14, pp.70-80 (1999)

K.H.L. Ho, Learning Fuzzy Concepts by Example with Fuzzy Conceptual Graphs. In 1st Australian Conceptual Structures Workshop, 1994. Armidale, Australia.

L. Hoebel and C. Welty (1999) Garbage collection, Intelligence (ACM) 10(2) : 48.

J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", Proceedings of the National Academy of Sciences U.S.A., Vol.79, pp.2554-2558 (1982)

J. J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons, Proceedings of the National Academy of Sciences U.S.A., Vol.81, pp.3088-3092 (1984)

A. Joshi and R. Krishnapuram, Robust Fuzzy Clustering Methods to Support Web Mining, in Proc Workshop in Data Mining and Knowledge Discovery, SIGMOD, pp. 15-1 to 15-8, 1998.

B. Kao, J. Lee, C. Y. Ng and D. Cheung (2000) Anchor point indexing in Web document retrieval, IEEE trans. on SMC (part C) 30(3) : 364-373.

J. Kacprzyk and A. Ziolkowski, Retrieval from databases using queries with fuzzy linguistic quantifiers, Fuzzy logic in knowledge engineering (Edited by Prade H and Negoita C. V), Verlag TUV Rheinland, 1986.

M. Kobayashi, K. Takeda, "Information retrieval on the web", ACM Computing Survey, Vol.32, pp.144-173 (2000)

B. Kosko, "Adaptive Bi-directional Associative Memories," Applied Optics, Vol. 26, No. 23, 4947-4960 (1987).

B. Kosko, "Neural Network and Fuzzy Systems," Prentice Hall (1992).

D. H. Kraft and F. E. Petry (1997) Fuzzy Information systems : managing uncertainty in databases and information retrieval systems, Fuzzy sets and systems 90(2) : 183-191.

R. Krishnapuram et al., A Fuzzy Relative of the K-medoids Algorithm with application to document and Snippet Clustering , in Proceedings of IEEE Intel. Conf. Fuzzy Systems-FUZZIEEE 99, Korea, 1999.

T. B. Lee , Transcript of Tim Berners-Lee's talk to the LCS 35th Anniversary celebrations, Cambridge Massachusetts, 1999/April/14

D. B. Lenat, From 2001 to 2001: Common Sense and the Mind of HAL; A chapter from Hal's Legacy: 2001 as Dream and Reality (http://www.cyc.com/publications.html)

Lidsky D and Kwon R (1997) Searching the net, PC magazine Dec. 2 : 227-258.

T. P. Martin, Searching and smushing on the Semantic Web – Challenges for Soft Computing, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

J. McCarthy (2000) Phenomenal data mining, Communications of the ACM 43(8) : 75-79.

H. Nakajima, T. Sogoh and M. Arao (1993) Development of an efficient fuzzy SQL for a large scale fuzzy relational database, Proc. 5th IFSA world congress : 517-530.

M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

M. Nikravesh, Fuzzy Logic and Internet: Perception Based Information Processing and Retrieval, Berkeley Initiative in Soft Computing, Report No. 2001-2-SI-BT, September 2001a.

M. Nikravesh, BISC and The New Millennium, Perception-based Information Processing, Berkeley Initiative in Soft Computing, Report No. 2001-1-SI, September 2001b.

S. K. Pal, V. Talwar, and P. Mitra, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, to be published in IEEE Transcations on Neural Networks, 2002.

F. Petry and P. Bosc, Fuzzy databases : principles and applications, Kluwer, Norwell, MA, 1996.

G. Presser, Fuzzy Personalization, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

D. Rasmussen and R. R. Yager (1999) Finding fuzzy and gradual functional dependencies with summarySQL, Fuzzy sets and systems 106(2) : 131-142.

G. Salton, Automatic text processing : the transformation, analysis and retrieval of information by computer, Addison-Wesley, Reading, MA, 1989.

G. Salton and M. J. McGill (1983) Introduction to modern information retrieval, McGraw-Hill.

E. Sanchez, Fuzzy logic e-motion, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

A. M. G. Serrano, Dialogue-based Approach to Intelligent Assistance on the Web, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

S. Shahrestani, Fuzzy Logic and Network Intrusion Detection, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

T. Takagi and M. Tajima, Proposal of a Search Engine based on Conceptual Matching of Text Notes, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

T. Takagi, A. Imura, H. Ushida, and T. Yamaguchi, "Conceptual Fuzzy Sets as a Meaning Representation and their Inductive Construction," International Journal of Intelligent Systems, Vol. 10, 929-945 (1995).

T. Takagi, A. Imura, H. Ushida, and T. Yamaguchi, "Multilayered Reasoning by Means of Conceptual Fuzzy Sets," International Journal of Intelligent Systems, Vol. 11, 97-111 (1996).

T. Takagi, S. Kasuya, M. Mukaidono, T. Yamaguchi, and T. Kokubo, "Realization of Sound-scape Agent by the Fusion of Conceptual Fuzzy Sets and Ontology," 8th International Conference on Fuzzy Systems FUZZ-IEEE'99, II, 801-806 (1999).

T. Takagi, S. Kasuya, M. Mukaidono, and T. Yamaguchi, "Conceptual Matching and its Applications to Selection of TV Programs and BGMs," IEEE International Conference on Systems, Man, and Cybernetics SMC'99, III, 269-273 (1999).

C. A. Testemale, Database system dealing with incomplete or uncertain information and vague queries, Fuzzy logic in knowledge engineering (Edited by Prade H and Negoita CV), Verlag TUV Rheinland, 1986.

M. Williams (1984) What makes rabbit run ?, J. man-mach. Stud. 2a (1) : 333-352.

Wittgenstein, "Philosophical Investigations," Basil Blackwell, Oxford (1953).

R. Yager, Aggregation Methods for Intelligent Search and Information Fusion, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

R. Yager (1991) On linguistic summaries of data, In knowledge discovery in databases, Piatetsky-Shapiro G and Frawley B (Eds.), MIT Press : 347-363.

R. Yager (1996) Database discovery using fuzzy sets, Int. J. of. Intelligence systems 11: 691-712.

J. Yen, Incorporating Fuzzy Ontology of Terms Relations in a Search Engine, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

L. A.Zadeh, The problem of deduction in an environment of imprecision, uncertainty, and partial truth, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001 [2001a].

L.A. Zadeh, A Prototype-Centered Approach to Adding Deduction Capability to Search Engines -- The Concept of Protoform, BISC Seminar, Feb 7, 2002, UC Berkeley, 2002.

L. A. Zadeh, " A new direction in AI – Toward a computational theory of perceptions, AI Magazine 22(1): Spring 2001, 73-84

L. A. Zadeh, From Computing with Numbers to Computing with Words-From Manipulation of Measurements to Manipulation of Perceptions, IEEE Trans. On Circuit and Systems-I Fundamental Theory and Applications, 45(1), Jan 1999, 105-119.

L. A. Zadeh (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning
and fuzzy logic, Fuzzy sets and systems 90(2) : 111-127.

L. A. Zadeh (1999) From computing with numbers to computing with words – From manipulation of measurements to manipulation of perceptions, IEEE trans. on circuit and systems 45(1) : 105-119.

L. A. Zadeh (1983) A computational approach to fuzzy quantifiers in natural language, Co put. Math.Appl. 9 : 149-184.

Y. Zhang et al., Granular Fuzzy Web Search Agents, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

Y. Zhang et al., Fuzzy Neural Web Agents for Stock Prediction, in M. Nikravesh and B. Azvine, FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28, August 2001.

Table 3. Understanding and History of Internet, World Wide Web and Search Engine;

Search Engine and Internet	Date	Developer	Affiliation	Comments
ARPANET	1962 –1969 1970-1973 1974-1981	UCLA	Under Leadership of DARPA	Initially designed to keep military sites in communication across the US. In 1969, ARPANET connected researchers from Stanford University, UCLA, UC Santa Barbara and the University of Utah. Internet community formed (1972). Email started (1977).
ALOHANET	1970		University of Hawaii	
USENET	1979	Tom Truscott & Jim Ellis Steve Bellovin	Duke University & University of North Carolina	The first newsgroup.
ARPANET	1982-1987	Bob Kahn & Vint Cerf	DARPA & Stanford University	ARPANET became "Internet". Vinton Cerf "Father of the Internet". Email and Newsgroups used by many universities.
CERT	1988-1990	Computer Emergency Response Team		Internet tool for communication. Privacy and Security. Digital world formed. Internet worms & hackers. The World Wide Web is born.
Archie through FTP	1990	Alan Ematage	McGill University	Originally for access to files given exact address. Finally for searching the archive sites on FTP server, deposit and retrieve files.
Gopher	1991	A team led by Mark MaCahill	University of Minnesota	Gopher used to organize all kinds of information stored on universities servers, libraries, non-classified government sites, etc. Archie and Veronica, helped Gopher (Search utilities).
World Wide Web "alt.hypertext	1991	Tim Berners-Lee	CERN in Switzerland	The first World Wide Web computer code. "alt.hypertext." newsgroup with the ability to combine words, pictures, and sounds on Web pages
Hyper Text Transfer Protocol (HTTP).	1991	Tim Berners-Lee	CERN in Switzerland	The 1990s marked the beginning of World Wide Web which in turn relies on HTML and Hyper HTTP. Conceived in 1989 at the CERN Physics Laboratory in Geneva. The first demonstration December 1990. On May 17, 1991, the World Wide Web was officially started, by granting HTTP access to a number of central CERN computers. Browser software became available-Microsoft Windows and

				Apple Macintosh
	1992			The first audio and video broadcasts:"MBONE." More than 1,000,000 hosts.
Veronica	1993	System Computing Services Group	University of Nevada	The search Device was similar to Archie but search Gopher servers for Text Files
Mosaic	1993	Marc Andeerssen	NCSA (the National Center for Supercomputing Applications); University of Illinois at Urbana Champaign	Mosaic, Graphical browser for the World Wide Web, were developed for the Xwindows/UNIX, Mac and Windows.
World Wide Web Wanderer; the first Spider robot	1993	Matthew Gary	MIT	Developed to count the web servers. Modified to capture URLs. First searchable Web database, the Wandex.
ALIWEB	1993	Martiijn Koster	Now with Excite	Archie-Like Indexing of the Web. The first META tag
JumpStation, World Wide Web Worm .	1993		NASA	Jump Station developed to gathere document titles and headings. Index the information by searching database and matching keywords. WWW worm index title tags and URLs.
Repository-Based Software Engineering (RBSE) Spider	1993		NASA	The first relevancy algorithm in search results, based on keyword frequency in the document. Robot-Driven Search Engine Spidered by content.
	1994			Broadcast over the M-Bone. Japan's Prime Minister goes online at www.kantei.go.jp. Backbone traffic exceeds 10 trillion bytes per month.
Netscape and Microsoft's Internet Explorer	1994-1998	Microsoft and Netscape	Microsoft and Netscape	Added a user-friendly point-and-click interface for browsing

Netscape	1994	Dr. James H. Clark and Marc Andreessen		The company was founded in April 1994 by Dr. James H. Clark, founder of Silicon Graphics, Inc. and Marc Andreessen, creator of the NCSA Mosaic research prototype for the Internet. June 5, 1995 - change the character of the World Wide Web from static pages to dynamic, interactive multimedia.
Galaxy	1994	Administered by Microelectronics and computer Technology Corporation	Funded by DARPA and consortium of technologies companies and original prototype by MADE program.	Provided large-scale support for electronic commerce and links documents into hierarchical categories with subcategories. Galaxy merged into Fox/News in 1999.
WebCrawler	1994	Brian Pinkerton	University of Washington	Search text of the sites and used for finding information in the Web. AOL purchased WebCrawler in 1995. Excite purchased WebCrawler in 1996.
Yahoo!	1994	David Filo and Jerry Yang	Stanford University	Organized the data into searchable directory based on simple database search engine. With the addition of the Google, Yahoo! Is the top-referring site for searches on the Web. It led also the future of the internet by changing the focus from search retrieval methods to clearly match the user's intent with the database.
Lycous	1994	Michael Mauldin	Carnegie Mellon University	New features such as ranked relevance retrieval, prefix matching, and word proximity matching. Until June 2000, it had used Inktomi as its back-end database provide. Currently, FAST a Norwegian search provider, replaced the Inktomi.
Excite	1995	Mark Van Haren, Ryan McIntyre, Ben Lutch, Joe Kraus, Graham Spencer, and Martin Reinfried	Architext Sofware	Combined search and retrieval with automatic hypertext linking to document and includes subject grouping and automatic abstract algorithm. IT can electronically parse and abstract from the web.
Infoseek	1995	Steve Kirsch	Infoseek	Infoseek combined many functional elements seen in other search

		(now with Propel)		tools such as Yahoo! And Lycos, but it boasted a solid user-friendly interface and consumer-focused features such as news. Also speed in which indexed Web sites and then added them to its live search database.
AltaVista	1995	Louis Monier, with Mike Burrows	Digital Equipment Corporation	Speed and the first "Natural Language" queries and Boolean operators. It also proved a user-friendly interface and the first search engine to add a link to helpful search tips below search field to assist novice searchers.
MetaCrawler	1995	Erick Selberg and Oren Etizinoi	University of Washington	The first Meta search engine. Search several search engines and reformat the results into a single page.
SavvySearch	1995	Daniel Dreilinger	Colorado State University	Meta Search which was included 20 search engines. Today, it includes 200 search engine.
Inktomi and HotBot	1994-1996	Eric Brewer and Paul Gauthier	University of California-Berkeley Funded by ARPA	Cluster inexpensive workstation computers to achieve the same computing power as expensive super computer. Powerful search technologies that made use of the clustering of workstations to achieve scaleable and flexible information retrieval system. HotBot, powered by Inktomi and was able to rapidly index and spider the Web and developing a very large database within a very short time.
LookSmart	1996	Mr Evan Thornley	LookSmart	Delivers a set of categorized listing presented in a user-friendly format and providing search infrastructure for vertical portals and ISPs.
AskJeeves	1997	Davis Warthen and Garrett Gruener	AskJeeves	It is built based on a large knowledge base on pre-searched Web sites. It used sophisticated, natural-language semantic and syntactic processing to understand the meaning of the user's question and match it to a 'question template" in the knowledge base.
GoTo	1997	Bill Gross	Indealab!	Auctioning off search engine positions. Advertisers to attach a value to their search engine placement.

Snap	1997	Halsey Minor, CNET Founder	CNET, Computer Network	Redefining the search engine space with a new business model; "portal" as first partnership between a traditional media company and an Internet portal.
Google	1997-1998	Larry Page and Sergey Brin	Stanford University	PageRank™ to deliver highly relevant search results based on proximity match and link popularity algorithms. Google represent the next generation of search engines.
Northern Light	1997	Team of librarians, software engineers, and information industry	Northern Light	To Index and classify human knowledge and has two database 1) contains an index to the full text of millions of Web pages and 2) includes full-text articles from a variety of sources. It searches both Web pages and full-text articles and sorts its search results into folders based on keywords, source, and other criteria.
AOL, MSN and Netscape	1998	AOL, MSN and Netscape	AOL, MSN and Netscape	Search service for the users of services and software
Open Directory	1998	Rick Skrenta and Bob Truel	dmoz	Open directory
Direct Hit	1998	Mike Cassidy	MIT	Direct Hit is dedicated to providing highly relevant Internet search results. Direct Hit's highly scalable search system leverages the searching activity of millions of Internet searchers to provide dramatically superior search results. By analyzing previous Internet search activity, Direct Hit determines the most relevant sites for your search request.
FAST Search	1999	Isaac Elsevier	FAST; Norwegian Company- All the Web	High-capacity search and real-time content matching engines based on the All the Web technology. Using Spider technology to index pages very rapidly. FAST can index both Audio and Video files.

Perception-Based Search Engines

Lotfi A. Zadeh
Berkeley Initiative in Soft Computing (BISC)
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
CA 94720-1776;
Zadeh@eecs.berkeley.edu
Telephone: 510-642-4959; Fax: 510-642-1712

From Search Engines to Question-Answering Systems: The Need for New Tools

Search engines, with Google at the top, have many remarkable capabilities. But what is not among them is the deduction capability—the capability to synthesize an answer to a query by drawing on bodies of information which are resident in various parts of the knowledge base. It is this capability that differentiates a question-answering system, Q/A system for short, from a search engine.

Construction of Q/A systems has a long history in AI. Interest in Q/A systems peaked in the seventies and eighties, and began to decline when it became obvious that the available tools were not adequate for construction of systems having significant question-answering capabilities. However, Q/A systems in the form of domain-restricted expert systems have proved to be of value, and are growing in versatility, visibility and importance.

Search engines as we know them today owe their existence and capabilities to the advent of the Web. A typical search engine is not designed to come up with answers to queries exemplified by "How many Ph.D. degrees in computer science were granted by Princeton University in 1996?" or "What is the name and affiliation of the leading eye surgeon in Boston?" or "What is the age of the oldest son of the President of Finland?" or "What is the fastest way of getting from Paris to London?"

Upgrading a search engine to a Q/A system is a complex, effort-intensive, open-ended problem. Semantic Web and related systems for upgrading quality of search may be viewed as steps in this direction. But what may be argued, as is done in the following, is that existing tools, based as they are on bivalent logic and probability theory, have intrinsic limitations. The principal obstacle is the nature of world knowledge.

The centrality of world knowledge in human cognition, and especially in reasoning and decision-making, has long been recognized in AI. The Cyc system of Douglas Lenat is a repository of world knowledge. The problem is that much of world knowledge consists of perceptions. Reflecting the bounded ability of sensory organs, and ultimately the brain, to resolve detail and store information, perceptions are intrinsically imprecise. More specifically, perceptions are f-granular in the sense that (a) the boundaries of perceived classes are fuzzy; and (b) the perceived values of attributes are granular, with a granule being a clump of values drawn together by indistinguishability, similarity, proximity or functionality. What is not widely recognized is that f-granularity of perceptions put them well beyond the reach of computational bivalent-logic-based theories. For example, the meaning of a simple perception described as "Most Swedes are tall," does not admit representation in predicate logic and/or probability theory.

Dealing with world knowledge needs new tools. A new tool which is suggested for this purpose is the fuzzy-logic-based method of computing with words and perceptions (CWP), with the understanding that perceptions are described in a natural language. A concept which plays a key role in CWP is that of Precisiated Natural Language (PNL). It is this language that is the centerpiece of our approach to reasoning and decision-making with world knowledge.

A concept which plays an essential role in PNL is that of precisiability. More specifically, a proposition, p, in a natural language, NL, is PL precisiable, or simply precisiable, if it is translatable into a mathematically well-defined language termed precisiation language, PL. Examples of precisiation languages are: the languages of propositional logic; predicate logic; modal logic; etc.; and Prolog; LISP; SQL; etc. These languages are based on bivalent logic. In the case of PNL, the precisiation language is a fuzzy-logic-based language referred to as the Generalized Constraint Language (GCL). By construction, GCL is maximally expressive.

A basic assumption underlying GCL is that, in general, the meaning of

a proposition, p, in NL may be represented as a generalized constraint of the form X isr R, where X is the constrained variable; R is the constraining relation, and r is a discrete-valued variable, termed modal variable, whose values define the modality of the constraint, that is, the way in which R constrains X. The principal modalities are; possibilistic (r=blank); probabilistic (r=p); veristic (r=v); usuality (r=u); fuzzy random set (r=rs); fuzzy graph (r=fg); and Pawlak set (r=ps). In general, X, R and r are implicit in p. Thus, precisiation of p, that is, translation of p into GCL, involves explicitation of X, R and r. GCL is generated by (a) combining generalized constraints; and (b) generalized constraint propagation, which is governed by the rules of inference in fuzzy logic. The translation of p expressed as a generalized constraint is referred to as the GC-form of p, GC(p). GC(p) may be viewed as a generalization of the concept of logical form. An abstraction of the GC-form is referred to as a protoform (prototypical form) of p, and is denoted as PF(p). For example, the protoform of p: "Most Swedes are tall" is Q A's are B's, where A and B are labels of fuzzy sets, and Q is a fuzzy quantifier. Two propositions p and q are said to be PF-equivalent if they have identical protoforms. For example, "Most Swedes are tall," and "Not many professors are rich," are PF-equivalent. In effect, a protoform of p is its deep semantic structure. The protoform language, PFL, consists of protoforms of elements of GCL.

With the concepts of GC-form and protoform in place, PNL may be defined as a subset of NL which is equipped with two dictionaries: (a) from NL to GCL; and (b) from GCL to PFL. In addition, PNL is equipped with a multiagent modular deduction database, DDB, which contains rules of deduction in PFL. A simple example of a rule of deduction in PFL which is identical to the compositional rule of inference in fuzzy logic, is: if X is A and (X, Y) is B then Y is A∘B, where A∘B is the composition of A and B, defined by $\mu_B(v) = \sup_u(\mu_A(u) \wedge \mu_B(u,v))$, where μ_A and μ_B are the membership functions of A and B, respectively, and $\wedge$ is min or, more generally, a T-norm. The rules of deduction in DDB are organized into modules and submodules, with each module and submodule associated with an agent who controls execution of rules of deduction and passing results of execution.

In our approach, PNL is employed in the main to represent information in the world knowledge database (WKD). For example, the items:

If X/Person works in Y/City then it is likely that X lives in or near Y
If X/Person lives in Y/City then it is likely that X works in or near Y

are translated into GCL as:

Distance (Location (Residence (X/Person), Location (Work (X/Person) isu near,

where isu, read as ezoo, is th e usuality constraint. The corresponding protoform is:

F (A(B(X/C), A(E(X/C)) isu G.

A concept which plays a key role in organization of world knowledge is that of an epistemic (knowledge-directed) lexicon (EL). Basically, an epistemic lexicon is a network of nodes and weighted links, with node i representing an object in the world knowledge database, and a weighted link from node i to node j representing the strength of association between i and j. The name of an object is a word or a composite word, e.g., car, passenger car or Ph.D. degree. An object is described by a relation or relations whose fields are attributes of the object. The values of an attribute may be granulated and associated with granulated probability and possibility distributions. For example, the values of a granular attribute may be labeled small, medium and large, and their probabilities may be described as low, high and low, respectively. Relations which are associated with an object serve as PNL-based descriptions of the world knowledge about the object. For example, a relation associated with an object labeled Ph.D. degree may contain attributes labeled Eligibility, Length.of.study, Granting.institution, etc. The knowledge associated with an object may be context-dependent. What should be stressed is that the concept of an epistemic lexicon is intended to be employed in representation of world knowledge — which is largely perception- based—rather than Web knowledge, which is not.

As a very simple illustration of the use of an epistemic lexicon, consider the query "How many horses received the Ph.D. degree from Princeton University in 1996." No existing search engine would come up with the correct answer, "Zero, since a horse cannot be a recipient of a Ph.D. degree." To generate the correct answer, the attribute Eligibility in the Ph.D. entry in EL should contain the condition "Human, usually over twenty years of age."

Conclusion

In conclusion, the main thrust of the fuzzy-logic-based approach to question-answering which is outlined in this abstract, is that to achieve significant question-answering capability it is necessary to develop methods of dealing with the reality that much of world knowledge—and especially knowledge about underlying probabilities is perception-based. Dealing with perception-based information is more complex and more effort-intensive than dealing with measurement-based information. In this instance, as in many others, complexity is the price that has to be paid to achieve superior performance.

Acknowledgements

Research supported in part by ONR N00014-00-1-0621, ONR Contract N00014-99-C-0298, NASA Contract NCC2-1006, NASA Grant NAC2-117, ONR Grant N00014-96-1-0556, ONR Grant FDN0014991035, ARO Grant DAAH 04-961-0341 and the BISC Program of UC Berkeley.

References

L. A. Zadeh, From Computing with Numbers to Computing with Words -- From Manipulation of Measurements to Manipulation of Perceptions, IEEE Transactions on Circuits and Systems, 45, 105-119, 1999.

L. A. Zadeh, "A new direction in AI: Towards a Computational Theory of Perceptions," *AI magazine*, vol. 22, pp. 73--84, 2001.

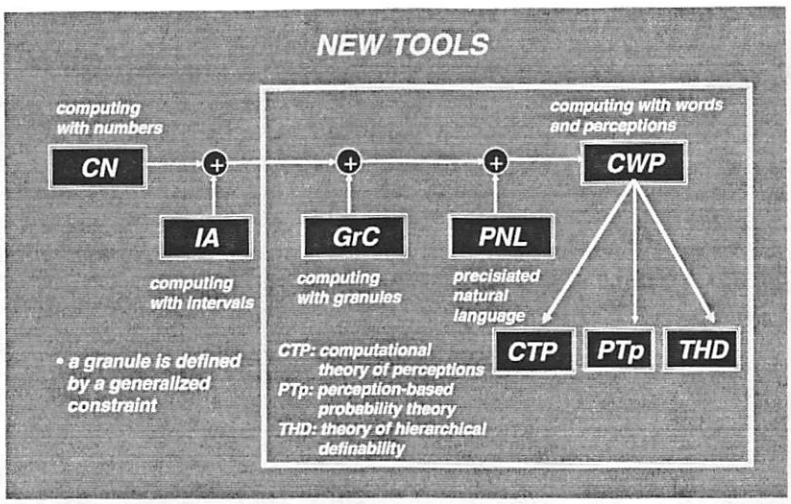
L.A. Zadeh, Toward a Perception-based Theory of Probabilistic Reasoning with Imprecise Probabilities, *Journal of Statistical Planning and Inference*, 105 233–264, 2002.

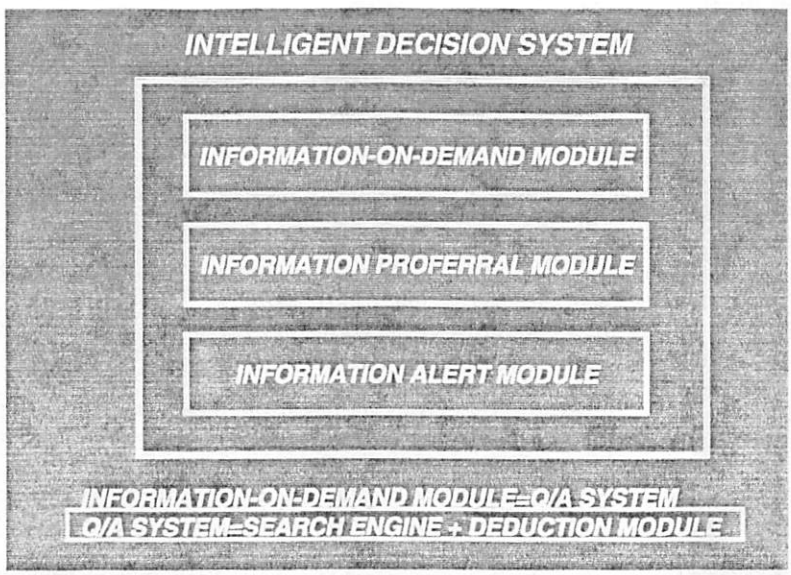
L. A. Zadeh and M. Nikravesh, Perception-Based Intelligent Decision Systems; Office of Naval Research, Summer 2002 Program Review, Covel Commons, University of California, Los Angeles, July 30th-August 1st, 2002.

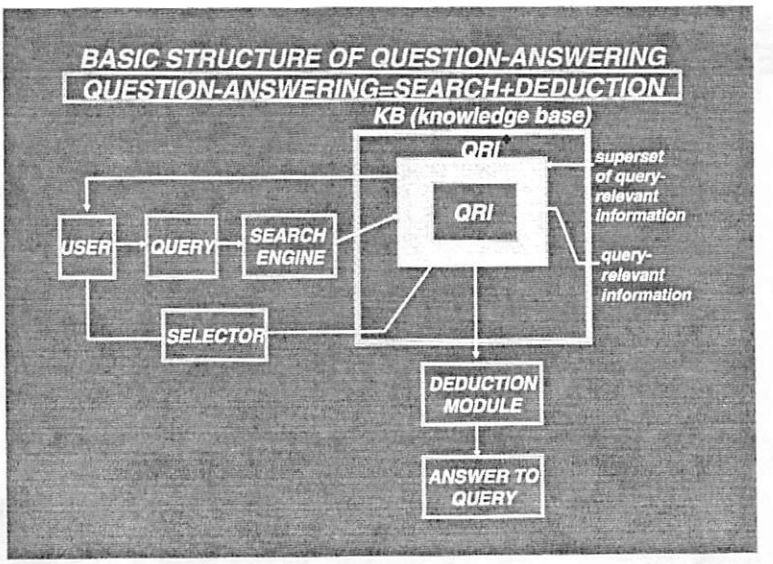
M. Nikravesh and B. Azvine: New Directions in Enhancing the Power of the Internet, Proc. Of the 2001 BISC Int. Workshop, University of California, Berkeley, Report: UCB/ERL M01/28, August 2001.

V. Loia , M. Nikravesh, L. A. Zadeh, *Journal of Soft Computing*, Special Issue; fuzzy Logic and the Internet, Springer Verlag, Vol. 6, No. 5; August 2002.

M. Nikravesh, R. Yager and L. A. Zadeh, "Enhancing the Power of the Internet", to be published in the Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer (August 2003).


NEW TOOLS
computing with numbers
CN
IA
computing with intervals
GrC
computing with granules
PNL
precisiated natural language
computing with words and perceptions
CWP
CTP
PTp
THD
• a granule is defined by a generalized constraint
CTP: computational theory of perceptions
PTp: perception-based probability theory
THD: theory of hierarchical definability


INTELLIGENT DECISION SYSTEM
INFORMATION-ON-DEMAND MODULE
INFORMATION PROFERRAL MODULE
INFORMATION ALERT MODULE
INFORMATION-ON-DEMAND MODULE=Q/A SYSTEM
Q/A SYSTEM=SEARCH ENGINE + DEDUCTION MODULE

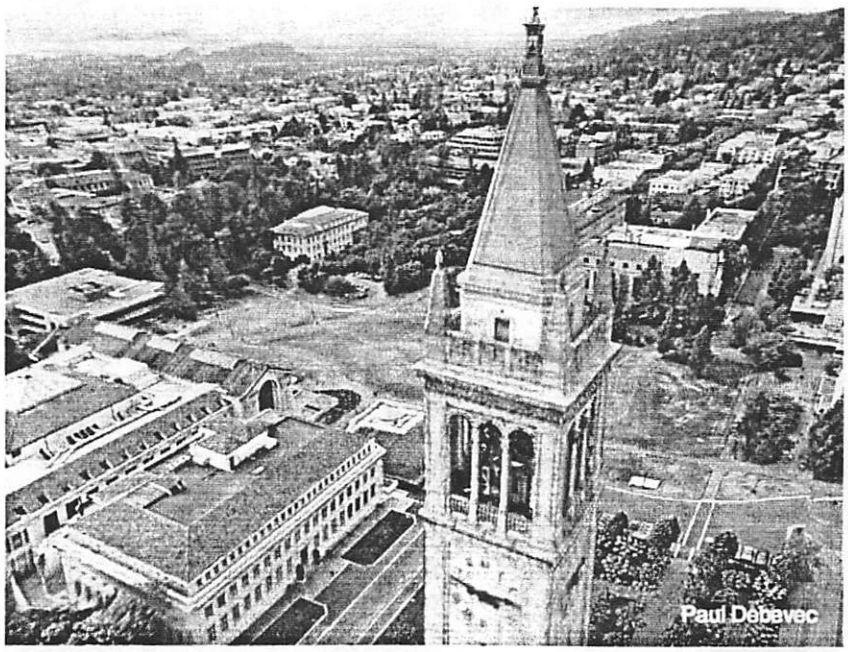

BASIC STRUCTURE OF QUESTION-ANSWERING
QUESTION-ANSWERING=SEARCH+DEDUCTION
KB (knowledge base)
QRI
QRI
superset of query-relevant information
query-relevant information
USER
QUERY
SEARCH ENGINE
SELECTOR
DEDUCTION MODULE
ANSWER TO QUERY

Fuzz-IEEE 2003

Special Track on Fuzzy Logic and the INTernet (FLINT)

ENHANCING THE POWER OF THE INTERNET

The IEEE International Conference on Fuzzy Systems
May 25-28, 2003
Marriott Pavilion Downtown Hotel
St. Louis, MO



Cal IEEE

Enhancing the Power of the Internet

Fuzz-IEEE 2003
Panel Session FlintPnl: Fuzzy Logic and the Internet
Monday, May 26, 4:30PM-6:00PM
Room: Ballroom: Salon D,
Chair: Masoud Nikravesh
Panelist: Lotfi A. Zadeh, Tomohiro Takagi, and Detlef Nauck

Short Description:

World Wide Web search engines have become the most heavily-used online services, with millions of searches performed each day. Their popularity is due, in part, to their ease of use. The central tasks for the most of the search engines can be summarize as 1) query or user information request- do what I mean and not what I say!, 2) model for the Internet, Web representation-web page collection, documents, text, images, music, etc, and 3) ranking or matching function-degree of relevance, recall, precision, similarity, etc.

Design of any new intelligent search engine should be at least based on two main motivations:

- The web environment is, for the most part, unstructured and imprecise and much of world knowledge consists of perceptions. To deal with information in the web environment what is needed is a logic that supports modes of reasoning which are approximate rather than exact. While searches may retrieve thousands of hits, finding decision-relevant and query-relevant information in an imprecise environment is a challenging problem, which has to be addressed. In addition, dealing with perception-based information is more complex and more effort intensive than dealing with measurement-based information. In this instance, as in many others, complexity is the price that has to be paid to achieve superior performance.

- Another, and less obvious, is deduction in an unstructured and imprecise environment given the huge stream of complex information. For Example; Search engines, with Google at the top, have many remarkable capabilities. But what is not among them is the deduction capability—the capability to synthesize an answer to a query by drawing on bodies of information which are resident in various parts of the knowledge base. It is this capability that differentiates a question-answering system, Q/A system for short, from a search engine. The main thrust of the fuzzy-logic-based approach to question-answering is that to achieve significant question-answering capability it is necessary to develop methods of dealing with the reality that much of world knowledge—and especially knowledge about underlying probabilities is perception-based.

From Search Engines to Question-Answering Systems The Need for New Tools

Lotfi A. Zadeh
Berkeley Initiative in Soft Computing (**BISC**)
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
CA 94720-1776;
Telephone: 510-642-4959; Fax: 510-642-1712
Zadeh@cs.berkeley.edu

Fuzz-IEEE 2003
FLINT Special Session

Abstract:

Search engines, with Google at the top, have many remarkable capabilities. But what is not among them is the deduction capability - the capability to synthesize an answer to a query by drawing on bodies of information which are resident in various parts of the knowledge base. It is this capability that differentiates a question-answering system, Q/A system for short, from a search engine. Upgrading a search engine to a Q/A system is a complex, effort-intensive, open-ended problem. Semantic Web and related systems for upgrading quality of search may be viewed as steps in this direction. But
what may be argued, as is done in the following, is that existing tools, based as they are on bivalent logic and probability theory, have intrinsic limitations. The principal obstacle is the nature of world knowledge.

The centrality of world knowledge in human cognition, and especially in reasoning and decision-making, has long been recognized in AI. The Cyc system of Douglas Lenat is a repository of world knowledge. The problem is that much of world knowledge consists of perceptions. More specifically, perceptions are f-granular in the sense that (a) the boundaries of perceived classes are fuzzy; and (b) the perceived values of attributes are granular, with a granule being a clump of values drawn together by indistinguishability, similarity, proximity or functionality. What is not widely recognized is that f-granularity of perceptions put them well beyond the reach of computational bivalent-logic-based
theories.

Dealing with world knowledge needs new tools. A new tool which is suggested for this purpose is the fuzzy-logic-based method of computing with words and perceptions (CWP), with the understanding that perceptions are described in a natural language. A concept which plays a key role in CWP is that of Precisiated Natural Language (PNL). It is this language that is the centerpiece of our approach to reasoning and decision-making with world knowledge.

A concept which plays a key role in organization of world knowledge is that of an epistemic (knowledge-directed) lexicon (EL). Basically, an epistemic lexicon is a network of nodes and weighted links, with node i representing an object in the world knowledge database, and a weighted link from node i to node j representing the strength of association between i and j. The name of an object is a word or a composite word, e.g., car, passenger car or Ph.D. degree. An object is described by a relation or relations whose fields are attributes of the object. The values of an attribute may be granulated and associated with granulated probability and possibility

distributions. For example, the values of a granular attribute may be labeled small, medium and large, and their probabilities may be described as low, high and low, respectively. Relations which are associated with an object serve as PNL-based descriptions of the world knowledge about the object. For example, a relation associated with an object labeled Ph.D. degree may contain attributes labeled Eligibility, Length.of.study, Granting.institution, etc. The knowledge associated with an object may be context-dependent. What should be stressed is that the concept of an epistemic lexicon is intended to be employed in representation of world knowledge - which is largely perception-based - rather than Web knowledge, which is not.

In conclusion, the main thrust of the fuzzy-logic-based approach to question-answering which is outlined in this abstract, is that to achieve significant question-answering capability it is necessary to develop methods of dealing with the reality that much of world knowledge is perception-based. Dealing with perception-based information is more complex and more effort-intensive than dealing with measurement-based information. In this instance, as in many others, complexity is the price that has to be paid to achieve superior performance.

Short Bio: Prof. Lotfi A. Zadeh; BISC Director

Prof. Zadeh is a Professor in the Graduate School,Computer Science Division, Department of EECS, University of California, Berkeley. In addition, he is serving as the Director of BISC (Berkeley Initiative in Soft Computing). His earlier work was concerned in the main with systems analysis, decision analysis and information systems. His current research is focused on fuzzy logic, computing with words and soft computing. Lotfi Zadeh is a Fellow of the IEEE, AAAS, ACM, AAAI, and IFSA. He is a member of the National Academy of Engineering and a Foreign Member of the Russian Academy of Natural Sciences. He is a recipient of the IEEE Education Medal, the IEEE Richard W. Hamming Medal, the IEEE Medal of Honor, the ASME Rufus Oldenburger Medal, the B. Bolzano Medal of the Czech Academy of Sciences, the Kampe de Feriet Medal, the AACC Richard E. Bellman Central Heritage Award, the Grigore Moisil Prize, the Honda Prize, the Okawa Prize, the AIM Information Science Award, the IEEE-SMC J. P. Wohl Career Acheivement Award, the SOFT Scientific Contribution Memorial Award of the Japan Society for Fuzzy Theory, the IEEE Millennium Medal, the ACM 2000 Allen Newell Award, and other awards and honorary doctorates.

Concept-Based Information Retrieval and Search Engine

Tomohiro Takagi
Department of Computer Science,
Meiji University
1-1-1 Higashi-Mita, Tama-ku,
Kawasaki-shi, Kanagawa-ken 214-8571 Japan
+81-44-934-7469
Takagi@cs.meiji.ac.jp

Fuzz-IEEE 2003
FLINT Special Session

Abstract

Since a fuzzy set is defined by enumerating its elements and the degree of membership of each element, we can use it to express word ambiguity by enumerating all possible meanings of a word, then estimating the degrees of compatibilities between the word and the meanings.

Based on this approach, we have proposed using conceptual fuzzy sets (CFSs) to represent the various meanings of a concept that change dynamically depending on the context. A CFS (is realized as neural networks in which a node represents a concept and a link represents the strength of the relation between two (connected) concepts. The activation values agreeing with the grades of membership are determined through this associative memory. In a CFS, the meaning of a concept is represented by the distribution of the activation values of the other nodes. The distribution evolves from the activation of the node representing the concept of interest.

This presentation will start with my motivation to propose CFSs and algorithm to generate CFSs. It will describe how it works to represent the context dependent meaning of a word and to measure a conceptual distance between documents. Next, information filtering and image search (Google-Based Search Engine for Multimedia Data) will be introduced as its applications to information retrieval using capability of conceptual matching. Finally we will introduce our approach to enhancing CFSs based on brain architecture.

Short Bio: Prof. Takagi received his B.Sc from Keio University and MSc. (Fuzzy Control and Reasoning) & PhD. (Fuzzy System Identification) in Computer Science from Tokyo Institute of Technology (1979 and 1983). Prof. Takagi currently is the Professor and also Chair of Computer Science Course in graduate school of Science and Technology of Meiji University from 2000 to 2001. From 1988-1998, he was the Manager, central research laboratory and corporate multimedia promotion division at Matsushita Electric Industrial Co., LTD. He was also the deputy director at the Laboratory for International Fuzzy Engineering Research (LIFE), which was a national project supported by the Ministry of International Trade and Industry, from 1991 to 1993. From 1984 to 1988, he was the Director, Development Division, Inter-field Systems Inc. From 1983 to 84, he was the EECS research fellow, Department of Electrical Engineering Computer Science, University of California Berkeley and in 1983 he received his Doctor of engineering degree from the Tokyo Institute of Technology. He Proposed the Takagi-Sugeno model, which is one of the most popular methodologies for developing fuzzy systems in the doctoral dissertation. Prof. Takagi has over 20 years research and industrial experience and worked as consultant to major companies and

funded several key projects in the area of soft computing. He published and presented over 100 articles on diverse topics and served as technical editor and several national and international technical committees and technical chairs including advisory board or technical review committee members for both government agencies and non-government agencies throughout the world. Prof. Takagi is the Member of IEEE, IEEE Computer Society, IEEE Communications Society, IEEE Systems Man & Cybernetics Society, and Association for Computing Machinery and Japan Society for Fuzzy Theory and Systems.

Computational Intelligence to Automate and Enhance the Intelligent Data Analysis Process

Detlef Nauck
BTexact Technologies
Adastral Park, United Kindom
+44.1473.605661
detlef.nauck@bt.com

Fuzz-IEEE 2003
FLINT Special Session

Abstract:

The computerization of all aspects of our daily live and the ever-growing use of the Internet make it ever easier to collect and store data. Nowadays customers expect that businesses cater for their individual needs. In order to personalize services, intelligent data analysis (IDA) and adaptive (learning) systems are required. Simple linear statistical analysis as it is mainly used in today's businesses cannot model complex dynamic dependencies that are hidden in the collected data. IDA goes one step further than today's data mining approaches and also considers the suitability of the created solutions in terms like usability, comprehension, simplicity and cost. The intelligence in IDA comes from the expert knowledge that can be integrated in the analysis process, the knowledge-based methods used for analysis and the new knowledge created and communicated by the analysis process.

In addition to statistical methods, today we also have modern intelligent algorithms based on computational intelligence and machine learning. Computational intelligent methods like neuro-fuzzy systems and probabilistic networks or AI methods like decision trees or inductive logic programming provide new, intelligent ways for analyzing data. All these methods are part of IDA. The advantage of IDA is that it both allows the inclusion of available knowledge and the extraction of new, comprehensible knowledge about the analyzed data.

In this talk I'll describe a platform for IDA that make extensive use of computational intelligence and soft computing to automate and enhance the IDA process. This platform - SPIDA - enables us to quickly derive and implement IDA solutions. Examples of such solutions

Short Bio: Dr. Detlef Nauck is working as a Chief Research Scientist in the Computational Intelligence Group of BTexact's Research Department, where he is currently leading a research program in Intelligent Data Analysis. Before that he worked as a Senior Researcher at the Department of Computer Science of the University of Braunschweig (1990-1996) and as a Senior Research Fellow and Senior Lecturer at the Faculty of Computer Science of the Otto-von-Guericke University of Magdeburg (1996-1999). He holds a Masters degree in Computer Science (1990) and a PhD in Computer Science (1994) both from the University of Braunschweig. He also holds a Venia Legendi in Computer Science (Habilitation) from the Otto-von-Guericke University of Magdeburg (2000). His research interests are in the area of Neural Networks and Fuzzy Systems, especially in their combinations, which are known as Neuro-Fuzzy Systems. He has developed several neuro-fuzzy learning algorithms that are able to derive linguistically interpretable rules from data. Since he has joined BTexact, Dr. Nauck has worked in several Intelligent Data

Analysis projects and in a project about creating autonomous machine learning systems. Dr. Nauck has published seven books and more than 70 papers and he is a regular member of program committees for conferences on computational intelligence. He is a Visiting Senior Lecturer at the University of Magdeburg and member of IEEE and the German Society of Computer Scientists (GI). Dr. Nauck is currently a member of the steering committee of EUNITE - the European Network of Excellence on Intelligent Technologies for Smart Adaptive Systems. EUNITE is funded by the Information Society Technologies Program (IST) within the European Union's Fifth RTD Framework Program. Dr. Nauck is the chairman of the EUNITE Research Committee on Integration of Intelligent Methods.

Web Intelligence: Conceptual Search Engine and Navigation

Masoud Nikravesh
Berkeley Initiative in Soft Computing (**BISC**)
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
CA 94720-1776;
Telephone: 510-643-4522; Fax: 510-642-5775
Nikravesh@cs.berkeley.edu

Fuzz-IEEE 2003
FLINT Special Session

Abstract:

In this presentation, first we will present the role of the fuzzy logic in the Internet. Then we will present an intelligent model that can mine the Internet to conceptually match and rank homepages based on predefined linguistic formulations and rules defined by experts or based on a set of known homepages. The FCM model will be used for intelligent information and knowledge retrieval through conceptual matching of both text and images (here defined as "Concept"). The FCM can also be used for constructing fuzzy ontology or terms related to the context of the query and search to resolve the ambiguity. This model can be used to calculate conceptually the degree of match to the object or query. We will also present the integration of our technology into commercial search engines such as Google ™ and Yahoo! as a framework that can be used to integrate our model into any other commercial search engines, or development of the next generation of search engines.

Short Bio: Dr. Nikravesh is the BISC Associate Director and BTExact technology Senior Fellow in the Computer Science Division, Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley and Research Scientist in the Imaging and Informatics Group at NERSC (National Energy Research Scientific Computing Division, Lawrence Berkeley National Laboratory. His credentials have led to front-page news at Lawrence Berkeley National Laboratory News and headline news at the Electronics Engineering Times. Dr. Nikravesh is the LBNL-NERSC (National Energy Research Scientific Computing Division) representative to the DiMI Executive Committee. Dr. Nikravesh has over 20 years research and industrial experience and worked as consultant to over 15 major companies and funded several key projects in the area of soft computing, data mining and fusion, control, and earth sciences through US government and major oil companies. He published and presented over 100 articles and published several books on diverse topics and served as technical editor and several national and international technical committees and technical chairs including advisory board or technical review committee members for both government agencies and non-government agencies throughout the USA and abroad. He served as member of IEEE, SPE, AICHE, SEG, AGU, and ACS.

Enhancing the Power of the Internet

Fuzz-IEEE 2003

Panel Session FIintPnl: Fuzzy Logic and the Internet
Monday, May 26, 4:30PM-6:00PM, Room: Ballroom: Salon D, Chair: Masoud Nikravesh
Panelist: Lotfi A. Zadeh, Tomohiro Takagi, and Detlef Nauck

Tuesday, May 27, 1:30PM-3:30PM, Room: Ballroom: Salon D, Chair: M. Nikravesh/O. Nasraoui

1:30PM
From search engines to question-answering systems: The need for new tools
Lotfi A. Zadeh

2:10PM
Concept-based web communities for Google search engine
Tomoe Tomiyama, Ryosuke Ohgaya, Akiyoshi Shinmura, Takayuki Kawabata, Tomohiro Takagi, and M. Nikravesh

2:30PM
Intention-aware information-delivery system
K. Hanamura, K. Kawabata, and Tomohiro Takagi

2:50PM
I-miner: A web usage mining framework using hierarchical intelligent systems
Ajith Abraham

3:10PM
An intelligent web recommendation engine based on fuzzy approximate reasoning
Olfa Nasraoui and Chris Petenes

Wednesday, May 28, 1:30PM-3:10PM, Room: Hotel Room: Salon A, Chair: M. Nikravesh/N. Mouaddib

1:30PM
A philosophical study on fuzzy sets and fuzzy applications
Tero T. Joronen

1:50PM
Fuzzy personalized wireless information agents
Yan-Qing Zhang, Wei Fan, and Jiannong Cao

2:10PM
Traffic engineering with MPLS using fuzzy logic for application in IP networks Raulison Alves Resende, Sandro M. Rossi, Akebo Yamakami, Luiz H. Bonani, and Edson Moschim

2:30PM
Improve TCP performance over ATM-UBR with FED+
Yoon-Tze Chin, Shiro Handa, Fumihito Sasamori, and Shinjiro Oshita

2:50PM
A fuzzy linguistic summarization technique for TV recommender systems
M. Gelgon, N. Mouaddib, A. Pigeau, G. Raschia, and R. Saint-Paul

Fuzz-IEEE 2003

Special Track on Fuzzy Logic and the INTernet (FLINT)

ENHANCING THE POWER OF THE INTERNET

The IEEE International Conference on Fuzzy Systems
May 25-28, 2003
Marriott Pavilion Downtown Hotel
St. Louis, MO

Cal IEEE