Focus

Soft Computing 6 (2002) 287 – 299 © Springer-Verlag 2002
DOI 10.1007/s00500-002-0182-x

Fuzzy logic and the Internet (FLINT): Internet, World Wide Web, and search engines

M. Nikravesh, V. Loia, B. Azvine

Abstract Retrieving relevant information is a crucial component of cased-based reasoning systems for Internet applications such as search engines. The task is to use user-defined queries to retrieve useful information according to certain measures. 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.

1
Introduction

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.

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.

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 Fig. 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 Fig. 1). 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 proba-

M. Nikravesh (✉)
BISC Program, EECS Department-CS Division,
University of California, Berkeley, CA 94720
E-mail: Nikravesh@cs.berkeley.edu

V. Loia
Dipartimento di Matematica e Informatica,
Università di Salerno,
84081 Baronissi (Salerno), Italy

B. Azvine
BTExact Technologies
Orion Building pp1/12, Adastral Park,
Martlesham, Ipswich IP5 3RE, UK

Funding for this research was provided by the British Telecommunication (BT) and the BISC Program of UC Berkeley.

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: +1720 Millions

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

bility 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 (Fig. 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. To improve the performance of a system one can use different mathematical model for aggregation operator for $(A \cap B)$ such as fuzzy logic. This will sift the curve to a higher value as is shown in Fig. 1b. 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 (Fig. 1b).

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 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.

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.

Lee (1999) in his transcript refers to the fuzzy concept and the human intuition with respect to the Web.

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, Zadeh's viewpoint (2001a, 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."

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

Table 3. (Contd.)

Search engine and Internet	Date	Developer	Affiliation	Comments
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	zSearch 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 (now with Propel)	Infoseek	Infoseek combined many functional elements seen in other search 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 in dexed 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

Table 4. Similarly/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: $\frac{|X \cap Y|}{\min(|X|.|Y|)}$
Disimilarity coefficient: $\frac{|X \Delta Y|}{|X|+|Y|} = 1 -$ dice's coefficient
$|X \Delta Y| = |X \cup Y| - |X \cap Y|$

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.

Lenat (2001) 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.

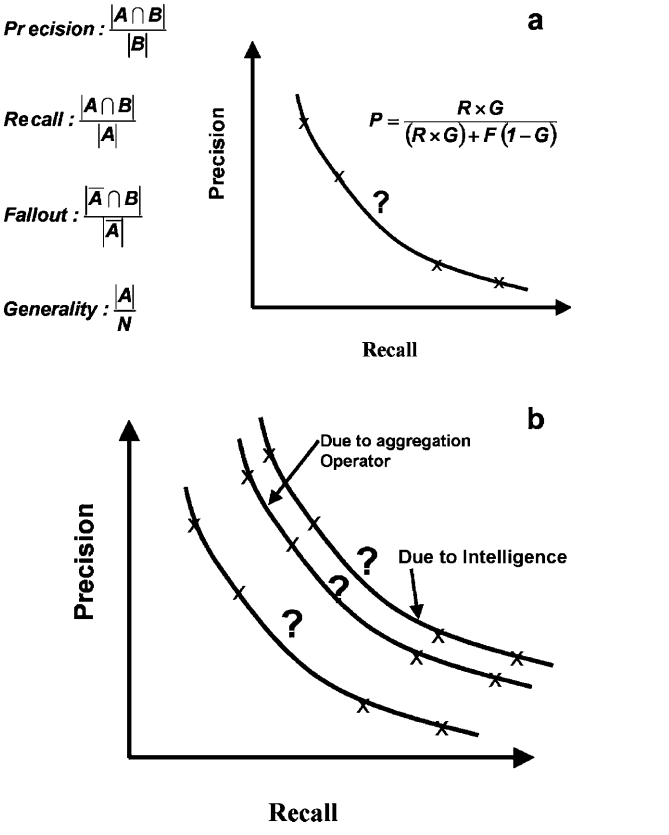

Fig. 1. **a** Relationship between precision and recall. **b** Effect of aggregation operator and Intelligent Techniques on Precision and Recall Performance

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 perception-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_1Q_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, 2001b; Nikravesh and Azvine, 2001; Nikravesh, 2001a, b) is one of the many ways that may help to address some of the issues presented by both Lee and 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

Table 6. Results of ranking in web search engines

Criteria	All terms		Proximity		Location	
First 20 and 100 items	20/100 hits (%)	Mean hits (%)	20/100 hits (%)	Mean hits (%)	20/100 hits (%)	Mean hits (%)
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

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
Alta Vista	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 power	X	X	*	*	*	X	*	*	*	*
Fast search advanced	X	X	*	*	*	X	*	*	*	*

* No information provided

computing with words (CW) (Zadeh, 1999). In CW, the objects of computation are words and propositions drawn from a natural language.

3
Fuzzy logic and the Internet

During the recent years, applications of fuzzy logic and the Internet from Web data mining to intelligent search engine and agents for Internet applications have greatly increased (Nikravesh and Azvine, 2001). Martin (2001) concluded that semantic web includes many aspects, which require fuzzy knowledge representation and reasoning. This includes the fuzzification and matching of concepts. In addition, it is concluded that fuzzy logic can be used in making useful, human-understandable, deduction from semi-structured information available in the web. It is also presented issues related to knowledge representation focusing on the process of fuzzy matching within graph structure. This includes knowledge representation based on conceptual graphs and Fril++. Baldwin and Morton (1985) studied the use of fuzzy logic in conceptual graph framework. Ho (1994) also used fuzzy conceptual graph to be implemented in the machine-learning framework. Baldwin (2001) presented the basic concept of fuzzy Bayesian Nets for user modeling, message filtering and data mining. For message filtering the protoype model representation has been used. Given a context, prototypes represent different types of people and can be modeled using fuzzy rules, fuzzy decision tree, fuzzy Bayesian Net or a fuzzy conceptual graph. In their study, fuzzy set has been used for better generalization. It has been also concluded that the new approach has many applications. For example, it can be used for personalization of web pages, intelligent filtering of the Emails, providing TV programs, books or movie and video of interest. Cao (2001) presented the fuzzy conceptual graphs for the semantic web. It is concluded that the use of conceptual graph and fuzzy logic is complementary for the semantic web. While conceptual graph provide a structure for natural language sentence, fuzzy logic provide a methodology for computing with words. It has been concluded that fuzzy conceptual graphs is suitable language for knowledge representation to be used by Semantic web. Takagi and Tajima (2001) presented the conceptual matching of text notes to be used by search engines. An new search engine

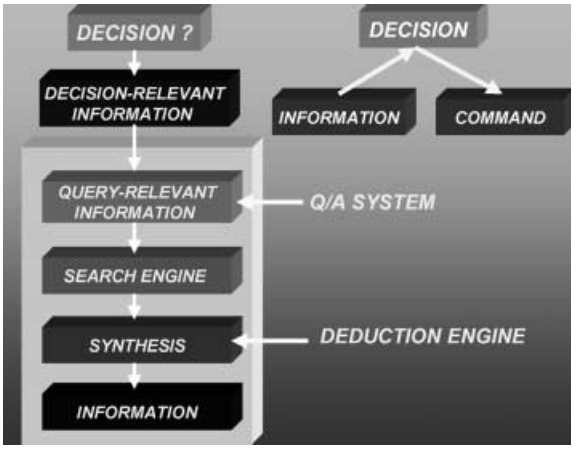

Fig. 2. Perception-based decision analysis (PDA) (Zadeh, 2001)

proposed which conceptually matches keywords and the web pages. Conceptual fuzzy set has been used for context-dependent keyword expansion. A new structure for search engine has been proposed which can resolve the context-dependent word ambiguity using fuzzy conceptual matching technique. Beremji (2001) used Fuzzy Reinforcement Learning (FRL) for text data mining and Internet search engine. Choi (2001) presented a new technique, which integrates document index with perception index. The techniques can be used for refinement of fuzzy queries on the Internet. It has been concluded that the use of perception index in commercial search engine provides a framework to handle fuzzy terms (perception-based), which is further step toward a human-friendly, natural language-based interface for the Internet. Sanchez (2001) presented the concept of Internet-based fuzzy Telerobotic for the WWW. The system receives the information from human and has the capability for fuzzy reasoning. It has been proposed to use fuzzy applets such as fuzzy logic propositions in the form of fuzzy rules that can be used for smart data base search. Bautista and Kraft (2001) presented an approach to use fuzzy logic for user profiling in Web retrieval applications. The technique can be used to expand the queries and knowledge extraction related to a group of users with common interest. Fuzzy representation of terms based on linguistic qualifiers has been used for their study. In addition, fuzzy clustering of the user profiles can be used to construct fuzzy rules and inferences in order to modify queries. The result can be used for knowledge extraction from user profiles for marketing purposes. Yager (2001) introduced fuzzy aggregation methods for intelligent search. It is concluded that the new technique can increase the expressiveness in the queries. Widyantoro and Yen (2001) proposed the use of fuzzy ontology in search engines. Fuzzy ontology of term relations can be built automatically from a collection of documents. The proposed fuzzy ontology can be used for query refinement and to suggest narrower and broader terms suggestions during user search activity. Presser (2001) introduced fuzzy logic for rule-based personalization and can be implemented for personalization of newsletters. It is concluded that the use of fuzzy logic provide better flexibility and better interpretation which helps in keeping the knowledge bases easy to maintain. Zhang et al. (2001a) presented granular fuzzy technique for web search engine to increase Internet search speed and the Internet quality of service. The techniques can be used for personalized fuzzy web search engine, the personalized granular web search agent. While current fuzzy search engines uses keywords, the proposed technique provide a framework to not only use traditional fuzzy-keyword but also fuzzy-user-preference-based search algorithm. It is concluded that the proposed model reduces web search redundancy, increase web search relevancy, and decrease user's web search time. Zhang et al. (2001b) proposed fuzzy neural web agents based on granular neural network, which discovers fuzzy rules for stock prediction. Fuzzy logic can be used for web mining. Pal et al. (2002) presented issues related to web mining using soft computing framework. The main tasks of web mining based on fuzzy logic include information retrieval and generalization. Krisnapuram et al. (1999) used fuzzy c medoids and triimed medoids for clustering of web documents. Joshi and Krisnapuram (1998) used fuzzy clustering for web log data mining. Sharestani (2001) presented the use of fuzzy logic for network intruder detection. It is concluded that fuzzy logic can be used for approximate reasoning and handling detection of intruders through approximate matching; fuzzy rule and summarizing the audit log data. Serrano (2001) presented a web-based intelligent assistance. The model is an agent-based system which uses a knowledge-based model of the e-business that provide advise to user through intelligent reasoning and dialogue evolution. The main advantage of this system is based on the human-computer understanding and expression capabilities, which generate the right information in the right time.

4 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 (Fig. 3).

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.

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

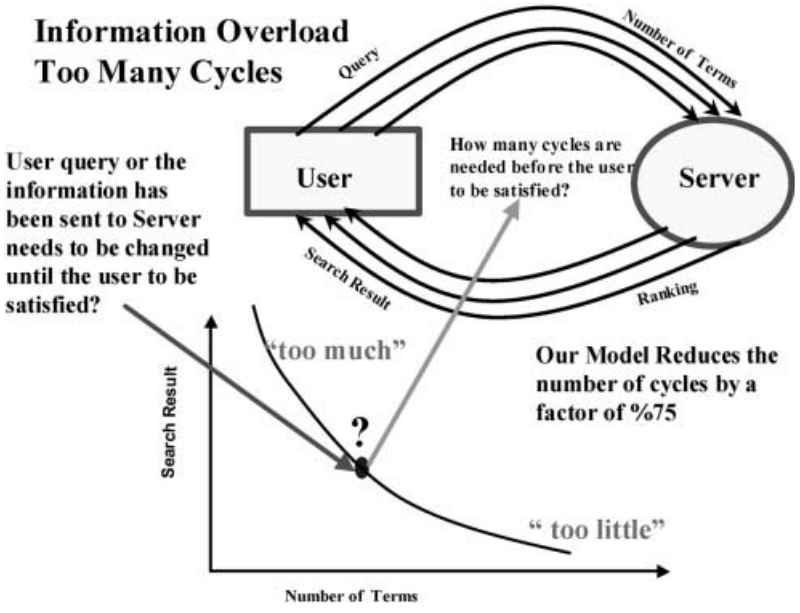

Fig. 3. Information overload

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*". Figure 4a and 4b 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.

5
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 (Fig. 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"). 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.

6
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-(FLINT) 2001, 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. Following are the areas that were recognized as challenging problems and the new direction toward the next generation of the search

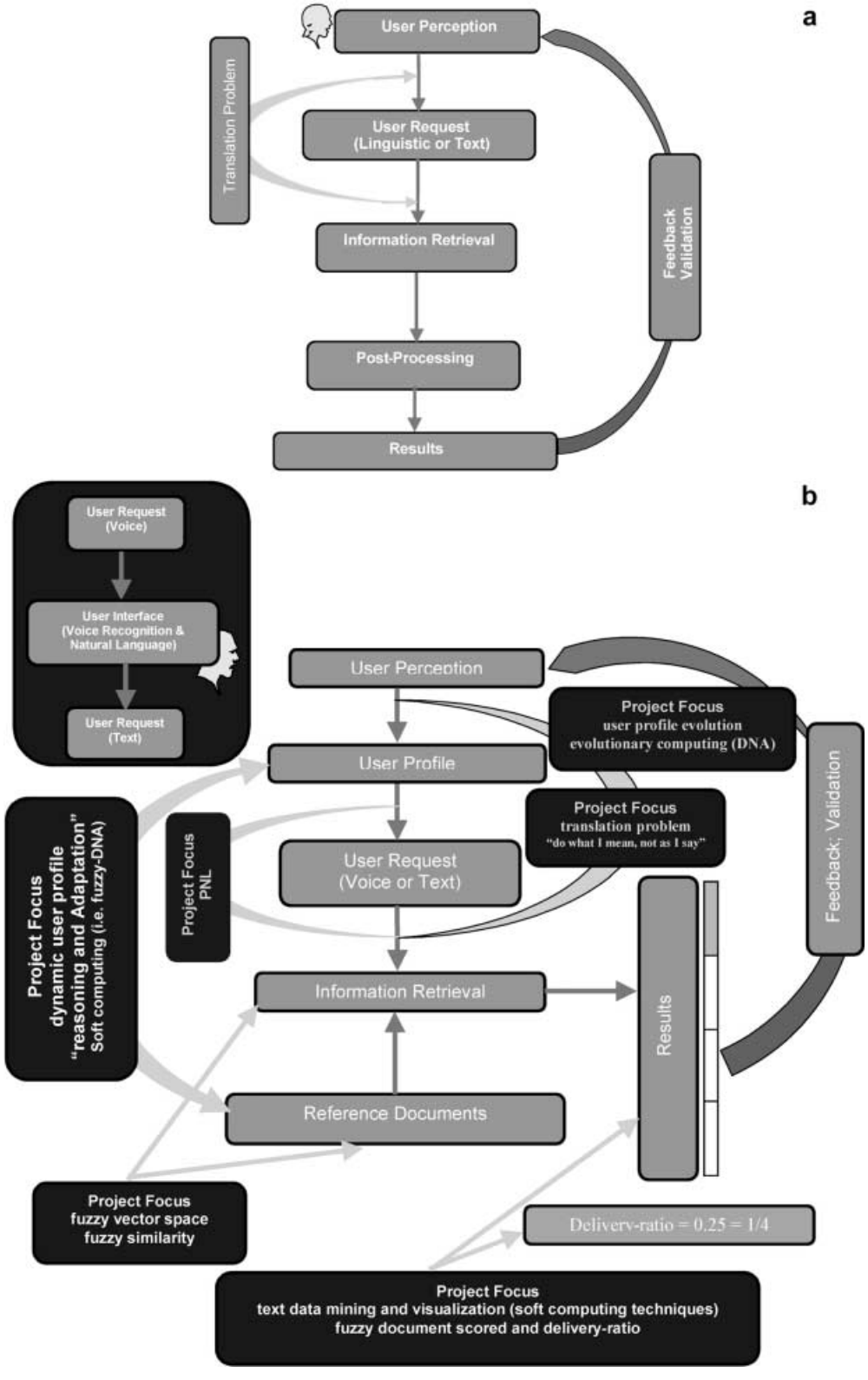

Fig. 4. **a** Structure of conventional search engine and retrieval technique. **b** Structure of search engine and retrieval technique and the problem related to perception and areas that soft computing can be used as a mean for improvement

engines and Internet. We summarize the challenges and the road ahead into four categories as follows:

1. Search engine and queries

- Deductive capabilities
- Customization and specialization
- Metadata and profiling
- Semantic web
- Imprecise-querying
- Automatic parallelism via database technology
- Approximate reasoning
- Ontology

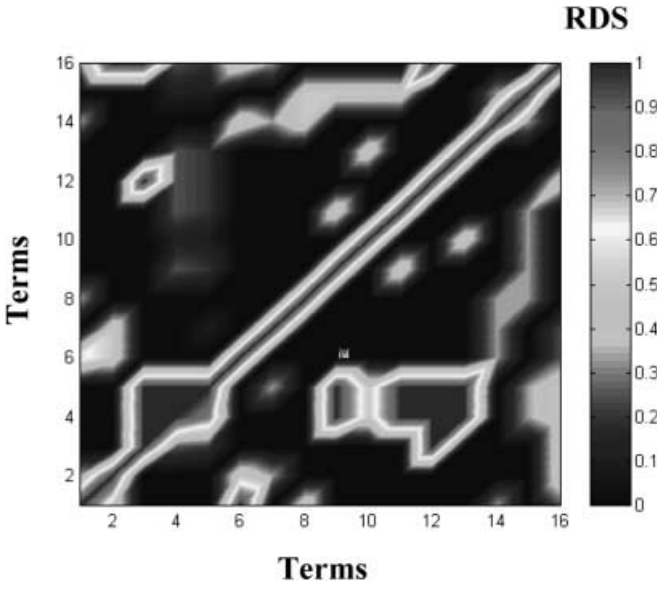

Fig. 5. Terms similarity; automated ontology generation and automated indexing

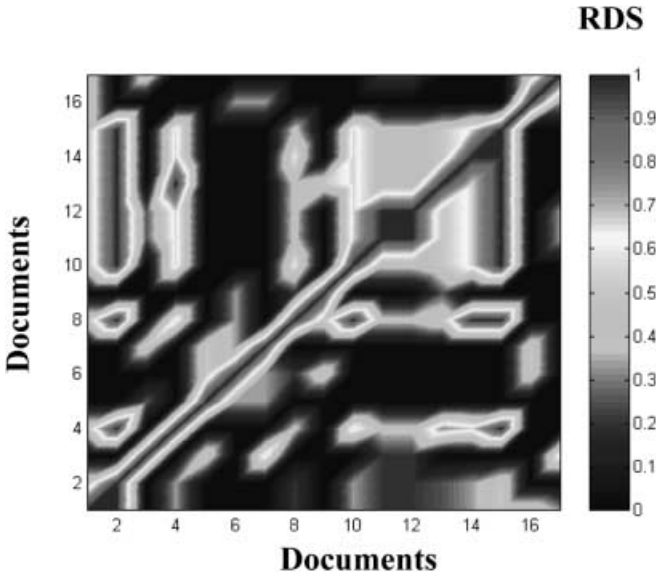

Fig. 6. Documents similarity; search personalization-user profiling

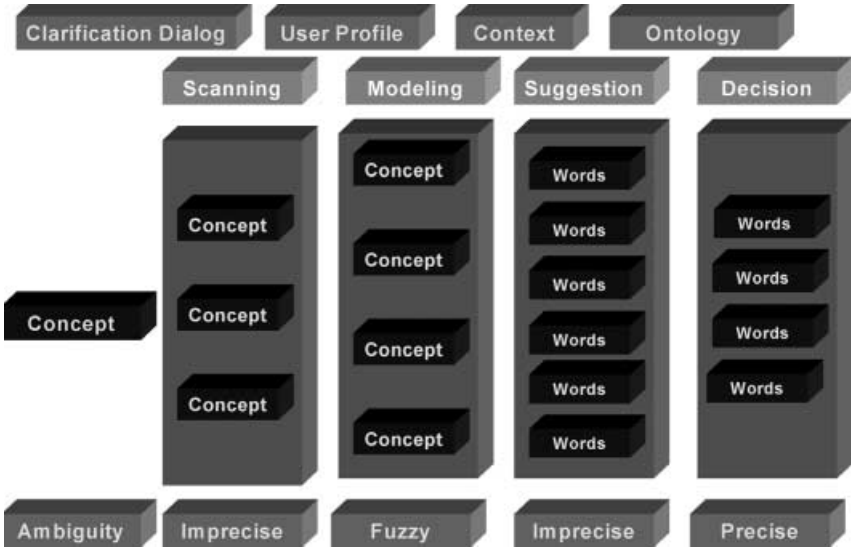

Fig. 7. Fuzzy conceptual matching and human mental model

- *Ambiguity resolution through clarification dialog; definition/meaning and specificity* user friendly
- Multimedia
- Databases
- Interaction

2. 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*)

3. 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

4. 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:

1. 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

2. 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.

7 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.

References

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

Baldwin JF, Morton SK (1985) Conceptual Graphs and Fuzzy Qualifiers in Natural Languages Interfaces, University of Bristol

Batista MJM et al (2001) User profiles and fuzzy logic in web retrieval. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Beremji H (2001) Fuzzy reinforcement learning and the internet with applications in power management or wireless networks. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Cao TH (2001) Fuzzy conceptual graphs for the semantic web. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Choi DY (2001) Integration of document index with perception index and its application to fuzzy query on the Internet. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Ho KHL (1994) Learning fuzzy concepts by example with fuzzy conceptual graphs. In 1st Australian Conceptual Structures Workshop, Armidale, Australia

Joshi A, Krishnapuram R (1998) Robust fuzzy clustering methods to support web mining. In Proc Workshop in Data Mining and Knowledge Discovery, SIGMOD, pp. 15-1 to 15-8

Krishnapuram R et al (1999) A fuzzy relative of the K-medoids algorithm with application to document and snippet clustering. In Proc IEEE Intel Conf Fuzzy Systems – FUZZIEEE 99, Korea

Lee TB (1999) Transcript of Tim Berners-Lee's talk to the LCS 35th Anniversary celebrations, Cambridge Massachusetts

Lenat DB (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)

Martin TP (2001) Searching and smushing on the semantic web – challenges for soft computing. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of

the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

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

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

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

Pal SK, Talwar V, Mitra P (2002) Web mining in soft computing framework: relevance, state of the art and future directions, IEEE Trans Neural Networks

Presser G (2001) Fuzzy personalization. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

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

Serrano AMG (2001) Dialogue-based approach to intelligent assistance on the web. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Shahrestani S (2001) Fuzzy logic and network intrusion detection. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Takagi T, Tajima M (2001) Proposal of a search engine based on conceptual matching of text notes. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Yager R (2001) Aggregation methods for intelligent search and information fusion. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Yen J (2001) Incorporating fuzzy ontology of terms relations in a search engine. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Zadeh LA (1999) From computing with numbers to computing with words-from manipulation of measurements to manipulation of perceptions, IEEE Trans Circuit and Systems-I Fundamental Theory and Applications **45**(1): 105–119

Zadeh LA (2001a) The problem of deduction in an environment of imprecision, uncertainty, and partial truth. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Zadeh LA (2001b) A new direction in AI – toward a computational theory of perceptions, AI Magazine **22**(1): 73–84

Zadeh LA (2002b) A Prototype-Centered Approach to Adding Deduction Capability to Search engines – the concept of protoform, BISC Seminar, UC Berkeley

Zhang Y et al (2001a) Granular fuzzy web search agents. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28

Zhang Y et al (2001b) Fuzzy neural web agents for stock prediction. In: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M01/28