WEB INTELLIGENCE:
CONCEPTUAL-BASED MODEL

by

Masoud Nikravesh, Tomohiro Takagi, Masanori Tajima,
Akiyoshi Shinmura, Ryosuke Ohgaya, Koji Taniguchi,
Kazuyosi Kawahara, Kouta Fukano and Akiko Aizawa

Memorandum No. UCB/ERL M03/19

5 June 2003

WEB INTELLIGENCE:
CONCEPTUAL-BASED MODEL

by

Masoud Nikravesh, Tomohiro Takagi, Masanori Tajima,
Akiyoshi Shinmura, Ryosuke Ohgaya, Koji Taniguchi,
Kazuyosi Kawahara, Kouta Fukano and Akiko Aizawa

Memorandum No. UCB/ERL M03/19

5 June 2003

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Web Intelligence: Conceptual-Based Model

Masoud Nikravesh[(1)], Tomohiro Takagi[(2)], Masanori Tajima[(2)], Akiyoshi Shinmura[(2)], Ryosuke Ohgaya[(2)], Koji Taniguchi[(2)], Kazuyosi Kawahara[(2)], Kouta Fukano[(2)], and Akiko Aizawa[(3)]

[(1)]BISC Program, EECS Department-CS Division
University of California, Berkeley, CA 94720
Nikravesh@cs.berkeley.edu
[(2)] Dept. of Computer Science, Meiji University.
[(3)]National Institute of Informatics

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. In this chapter, 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 Fuzzy Conceptual Matching (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.

1 Introduction

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 (Zadeh 2001a and 2002):

i· 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.

ii· Another, and less obvious, is deduction in an unstructured and imprecise environment given the huge stream of complex information.

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 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. Berenji (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 be 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-key-word 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.

In our perspective, one can use clarification dialog, user profile, context, and ontology, into an integrated frame work to design a more intelligent search engine. The model will be used for intelligent information and knowledge retrieval through conceptual matching of text. The selected query doesn't need to match the decision criteria exactly, which gives the system a more human-like behavior. The model can also be used for constructing ontology or terms related to the context of search or query to resolve the ambiguity. 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.

It is also possible to automate ontology generation and document indexing using the terms similarity based on Conceptual-Latent Semantic Indexing Technique (CLSI). 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. It is also possible to generate conceptual documents similarity map that can be used for intelligent search engine based on CLSI, 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.

Given the ambiguity and imprecision of the "concept" in the Internet, which may be described by both textual and image information, the use of Fuzzy Conceptual Matching (FCM) is a necessity for search engines. In the FCM approach, the "concept" is defined by a series of keywords with different weights depending on the importance of each keyword. Ambiguity in concepts can be defined by a set of imprecise concepts. Each imprecise concept in fact can be defined by a set of fuzzy concepts. The fuzzy concepts can then be related to a set of imprecise words given the context. Imprecise words can then be translated into precise words given the ontology and ambiguity resolution through clarification dialog. By constructing the ontology and fine-tuning the strength of links (weights), we could construct a fuzzy set to integrate piecewise the imprecise concepts and precise words to define the ambiguous concept.

Currently on the Internet there exists a host of illegal web sites which specialize in the distribution of commercial software and music. This chapter proposes a method to distinguish illegal web sites from legal ones not only by using tf-idf values but also to recognize the purpose/meaning of the web sites. It is achieved by describing what are considered to be illegal sites and by judging whether the objective web sites match the description of illegality. Conceptual fuzzy sets (CFSs) are used to describe the concept of illegal web sites. First, we introduced the usefulness of CFSs in overcoming those problems, and propose the realization of CFSs using RBF-like networks. In a CFS, the meaning of a concept is represented by the distribution of the activation values of the other nodes. Because the distribution changes depend on which labels are activated as a result of the conditions, the activations show a context-dependent meaning. Next, we proposed the architecture of the filtering system. Finally, we compared the proposed method with the tf-idf method with the support vector machine. The e-measures as a total evaluation indicate that the proposed system showed better results as compared to the tf-idf method with the support vector machine.

In addition, we propose a menu navigation system which conceptually matches input keywords and paths. For conceptual matching, we use conceptual fuzzy sets (CFSs) based on radial basis function (RBF) networks. In a CFS, the meaning of a concept is represented by the distribution of the activation values of the other concepts. To expand input keywords, the propagation of activation values is carried out recursively. The proposed system recommends users paths to appropriate categories. We use 3D user interface to navigate users.

2 Fuzzy Conceptual Model and Search Engine

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 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 workflow needs to be designed to make optimal use of this information. In this Chapter, we present the foundation of CFS-Based Intelligent Model and its applications to both information filtering and design of navigation.

2.1 Search Engine based on Conceptual Matching of Text Notes

Information retrieval in the Internet is generally done by using keyword matching, which requires that for words to match, they must be the same or synonyms. But essentially, not only the information that matches the keywords exactly, but also information related in meaning to the input keywords should be retrieved. The following reasons are why fuzzy sets are essential for information retrieval.

First, a fuzzy set is defined by enumerating its elements and the degree of membership of each element. It is useful for retrieving information which includes not only the keyword, but also elements of the fuzzy set labeled by the input keyword. For example, a search engine may use baseball, diving, skiing, etc., as kinds of sports, when a user inputs "sports" as the keyword.

Second, the same word can have various meanings. Several words are used concurrently in usual sentences, but each word has multiple possible meanings (region), so we suppose an appropriate context which suits all regions of meaning of all words (**Figure 1**). At the same time, the context determines the meaning of each word.

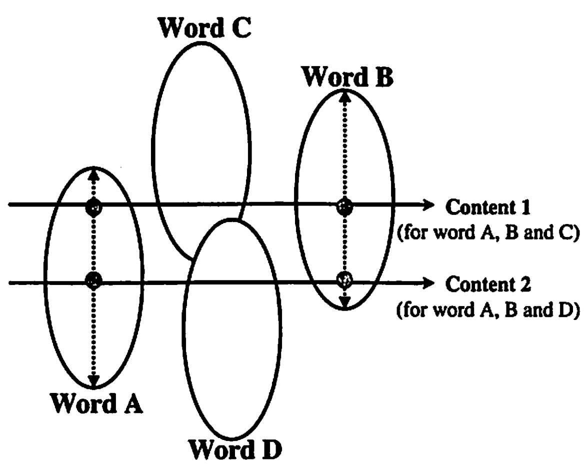


Figure 1. Meanings of words determined by a context.

For example, "sports" may mean "diving" or "sailing" when it is used with "marine," and may mean "baseball" or "basketball" when used with "TV programs." That is, each possibility distribution of meaning is considered as a fuzzy set itself. For information retrieval, keyword expansion that considers context is necessary, because simple expansion of a possible region causes a flood of words. For example, even if the user intends "marine sports," the set of expanded keywords includes unnecessary sports such as "baseball." However, an ordinary fuzzy set does not provide us the method to deal with context-dependent word ambiguity. To overcome this problem, we previously proposed using conceptual fuzzy sets (CFSs) (Takagi et al. 1995, 1996, 1999a and 1999b), which conform to Wittgenstein's concept, to represent the meanings of concepts.

In this section, we propose a search engine which conceptually matches input keywords and Web pages. The conceptual matching is attained by context dependent keyword expansion using conceptual fuzzy sets. We describe the necessity of conceptual fuzzy sets for information retrieval in Section 2.1.1, and propose the use of conceptual fuzzy sets using Hopfield Networks in section 2.1.2. Section 2.1.3 proposes the search engine which can execute conceptual matching and deal with context-dependent word ambiguity. In Section 2.1.4, we show two simulations of retrieving actual Web pages comparing the proposed method with the ordinary TF-IDF method. In section 2.1.5, we provide the summary.

2.1.1 Fuzzy Sets and Context Dependent Word Ambiguity

In this section we will present the context dependent word ambiguity and how to resolve the issue.

2.1.1.1 Conceptual Fuzzy Sets (Takagi et al. 1995, 1996, 1999a and 1999b)

Let's think about the meaning of "heavy." A person weighting 100kg would usually be considered heavy. But there is no clear boundary between "heavy" and "not heavy." Fuzzy sets are generally used to indicate these regions. That is, we have a problem of specificity.

Figure 2. The meaning of "heavy."

Let's think about it some more. For a vehicle, "heavy" might be several thousand kgs. For a ship, it might be more than ten thousand tons. Therefore, the item "heavy" being judged affects the vagueness of the meaning of "heavy" much more than the non-specificity of the amount when the item is already determined as shown in **Figure 2.** Moreover, "work" can be heavy, "traffic" can be heavy, and "smoking" can be heavy. So the meaning of "heavy" changes with the context, which results in the vagueness of the meaning.

That is, the main cause of vagueness is ambiguity in the language, not specificity. Ordinary fuzzy set theory has not dealt with the context dependent meaning representation concerning language ambiguity. However, as we mentioned in the Introduction, 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. Fuzzy set theory should therefore deal with language ambiguity as the main cause of vagueness.

To overcome this problem, we previously proposed using conceptual fuzzy sets. Although several works have been published, we will explain CFSs for understanding the following section. According to Wittgenstein (1953), the meaning of a concept can be represented by the totality of its uses. In this spirit, conceptual fuzzy sets, in which a concept is represented by the distribution of the activation concepts, are proposed.

The label of a fuzzy set represents the name of a concept, and a fuzzy set represents the meaning of the concept. Therefore, the shape of a fuzzy set is determined by the meaning of the label, which depends on the situation (Figure 3).

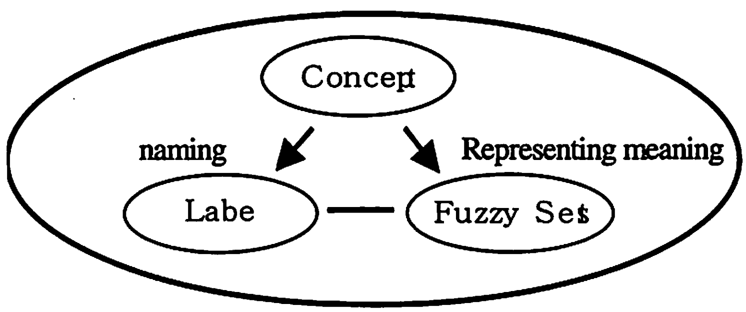


Figure 3. A fuzzy set as a meaning representation.

According to the theory of "meaning representation from use" proposed by Wittgenstein) the various meanings of a label (word) can be represented by other labels (words), and we can assign grades of activation showing the degree of compatibility between labels.

A CFS achieves this by using distributions of activations. A CFS is realized as an associative memory 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. The image of a CFS is shown in **Figure 4**.

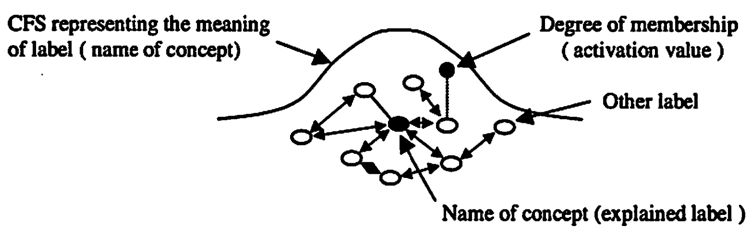


Figure 4. A conceptual fuzzy set represented by associative memories.

2.1.1.2 CFS Representing a Composed Concept having Multiple Meanings Depending on the Situation

Because the distribution changes depending on which labels are activated as a result of the conditions the activations show a context-dependent meaning. When more than two labels are activated, a CFS is realized by the overlapping propagations of their activations. In CFS notation, operations and their controls are all realized by the distribution of the activations and their propagations in the associative memories.

We can say that the distribution determined by the activation of a label agrees with the region of thought corresponding to the word expressing its meaning. The distribution (meaning of a label), that is a figure of a fuzzy set, changes depending on considered aspects that reflect the context.

2.1.2 Creating of Conceptual Fuzzy Sets

Previously we used bidirectional associative memories (BAMs) (Kasko 1987 and 1992) to generate CFSs, because of the clarity of the constraints used for their utilization. In this paper, we use Hopfield Networks, whose output can be also used with a continuous value, to overcome the limitation of BAMs that are a layered neural network. We do so because in a correlative neural network, relationships between concepts may not be a layered structure.

The following shows how to construct CFSs using Hopfield Networks (Hopfield 1982 and 1984).

Memorizing pieces of knowledge:

1. Classify piece of knowledge into several aspects. One piece becomes one pattern and one aspect corresponds to one Hopfield Network.
2. Each Hopfield network contains multiple patterns in the same aspect.

Generating CFSs:

1. Recollect patterns which include a node of interest in each Hopfield Network.
2. Sum all recollected patterns and normalize the activation values.

Figure 5 shows the image of memorized patterns and a generated CFS.

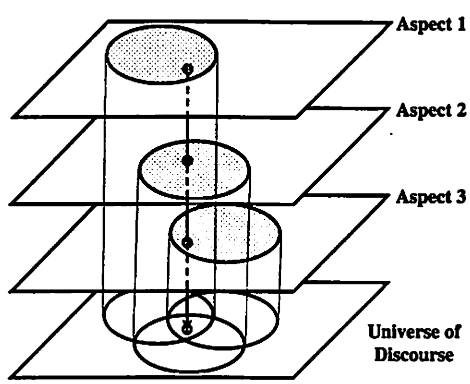


Figure 5. Image of memorized patterns and a generated CFS.

Logic-based paradigms for knowledge representation use symbolic processing both for concept representation and inference. Their underlying assumption is that a concept can be defined precisely. This causes an exponential increase in the number of cases, because we have to memorize all cases according to every possible context. In contrast, knowledge representation using CFSs memorizes knowledge about generalizations instead of exact cases (**Figure 5**). The following is an example to compare the proposed knowledge representation with ordinary logic-based paradigms. It shows that context-dependent meanings are generated by activating several keywords.

Example:

Let's think about the meaning of "heavy" again. The subject may be a human, a cat, or an elephant. Moreover, the age of the subject may be a baby or an adult, which also influences the meaning of "heavy." Therefore, since the number of cases increases exponentially as:

(cat, human, elephant......) *
(baby, adult, ...) * ....,

it is impossible to know how heavy the subjects are in all cases. On the other hand, using CFSs, which create meaning by overlapping activations, number of cases to be memorized becomes:

(cat, human, elephant......) +
(baby, adult.......) +...........,

and increases linearly.

Let's generate CFSs in these contexts. Assume the universe of discourse is "weight," from 0-1000 kg. Aspects and memorized patterns are as follows.

(Aspect)	(Memorized pattern)
kind	cat, human, elephant
age	baby, adult

<Step1> Memorize patterns such as those in **Figures 6** and **7** for each aspect. For example, the pattern "cat" shows that it memorizes its usual heavy weight within the activation range of [-1,1]. [-1,1] is the bi-polar expression of [0,1].

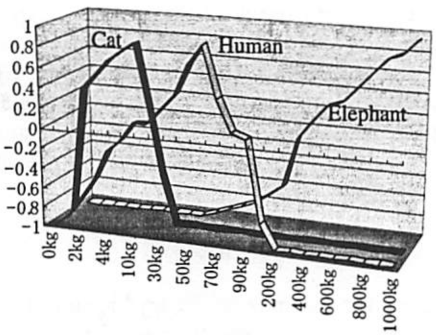


Figure 6. Memorized patterns of "heavy cat," "heavy human," and "heavy elephant"

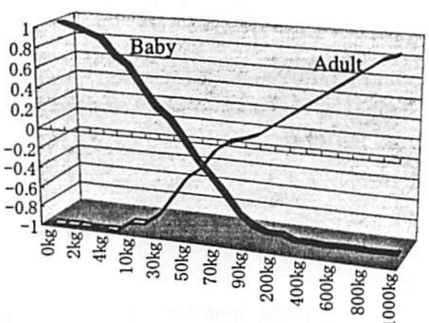


Figure 7. Memorized patterns of "heavy baby," and "heavy adult"

<Step2>When keywords are input, the input values of the neurons corresponding to these keywords are set as 1 and the input values of the other neurons are set as -1, and each Hopfield Network recollects the patterns.

<Step3> Finally, the activations of nodes in all aspects are summed up, and they are normalized in the range of [-1,1]. The normalized outputs become the final outputs result.

Figure 8 shows the ability of CFSs to generate context-dependent meanings of "heavy human" in the case of "adult" and "baby." We can recognize that both fuzzy sets have different shapes even when considering the same word "human." **Figure 9** compares the difference between the case of "human" and "elephant." Here, [a + b] means that the activation is started from the concepts in nodes "a "(ex: adult) and "b" (ex: elephant).

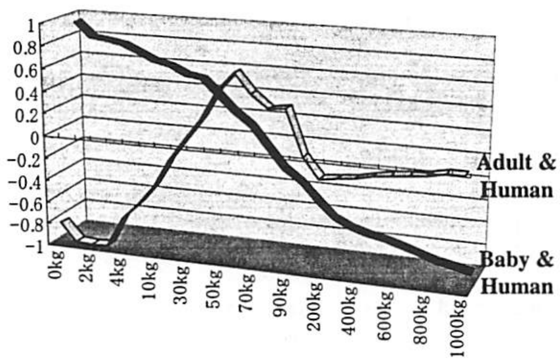


Figure 8. Example of outputs "heavy human."

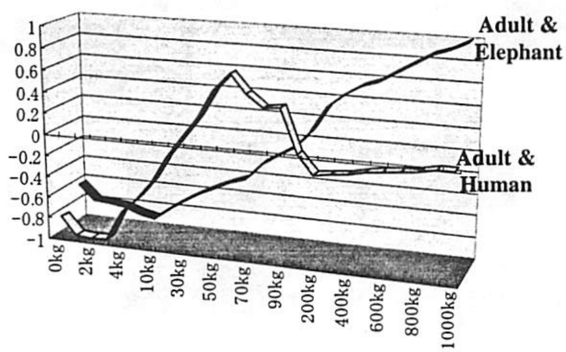


Figure 9. Example of outputs "heavy adult."

2.1.3 Conceptual Matching in a Search Engine Using CFS

In this section, we will focus on the use of CFS in search engines.

2.1.3.1 Scheme of Search Engine (Kobayashi and Takeda 2000, Quarino and Vetere 1999)

Usually, search engines work as follows: (We may want to add information about the search engines, my slides)

Index collecting of Web pages:

An indexer extracts words from Web pages, analyzes them, and stores the result as indexing information.

Retrieving information:

The Web pages, which include input keywords, are extracted. The pages are assigned priority and are sorted referencing the indexing information above.

As we mentioned earlier, information retrieval is generally done by using keyword matching, which requires words to match and is different from conceptual matching.

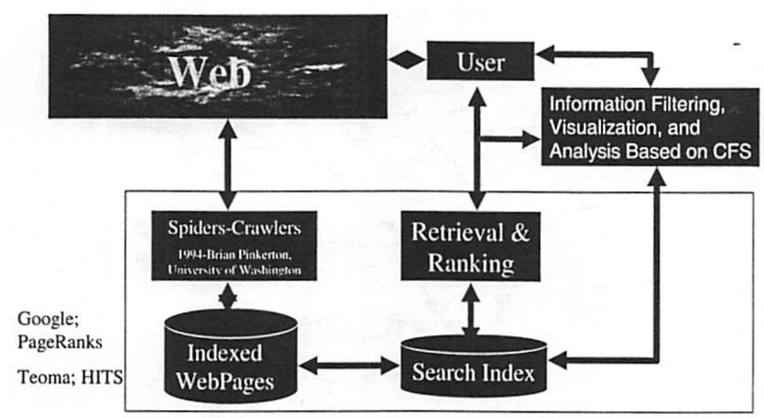


Figure 10. Search Engines Architecture

2.1.3.2 Conceptual Matching

We propose a search engine system which conceptually matches input keywords and Web pages according to the following idea.

1. Expand input keywords to the elements of CFSs.
2. Evaluate the matching degree between the set of expanded keywords and the set of words included in each Web page.
3. Sort the Web pages and display them according to the matching degrees.

The following shows the process in the proposed search engine.

Index collecting of Web pages:

1. Extract nouns and adjectives, and count the frequency of each word.
2. Calculate an evaluation of each word using the TF-IDF method for each Web page.
3. Store the evaluation into a lexicon.

Retrieve information:

1. A user inputs keywords into a browser, which transfers the keywords to a CFS unit.
2. Propagation of activation occurs from input keywords in the CFS unit. The meanings of the keywords are represented in other expanded words using conceptual fuzzy sets, and the activation value of each word is stored into the lexicon.
3. Matching is executed in the following process for each Web page. Obtain the final evaluation of each word by multiplying the evaluation by the TF-IDF method and the activation value. Sum up the final evaluations of all words and attach the result to each Web page as a matching degree.
4. The matched Web pages are sorted according to the matching degrees, and their addresses are returned to the browser with their matching degrees.

2.1.4 Simulations and Evaluations

Let's think about the case where we are searching for Web pages of places to visit using certain keywords, we indexed 200 actual Web pages, and compared the search result of the following two matching methods.

1. TF-IDF method
2. our proposed method (**Figure 11** using CFS)

Evaluation 1:

If the CFS unit has knowledge in fuzzy sets about places, and if a user inputs "famous resort" as a keyword, relating name of places are added as expanded keywords with their activation degrees agreeing with membership degrees.

Famous resort = 0.95/gold coast + 0.95/the Cote d'Azur + 0.91/Fiji + ..

Table 2.a shows the result when "famous resort" and "the Mediterranean Sea" are input as keywords. It consists of names of places and activation values, and shows the extended keywords generated by the activation of the above two keywords.

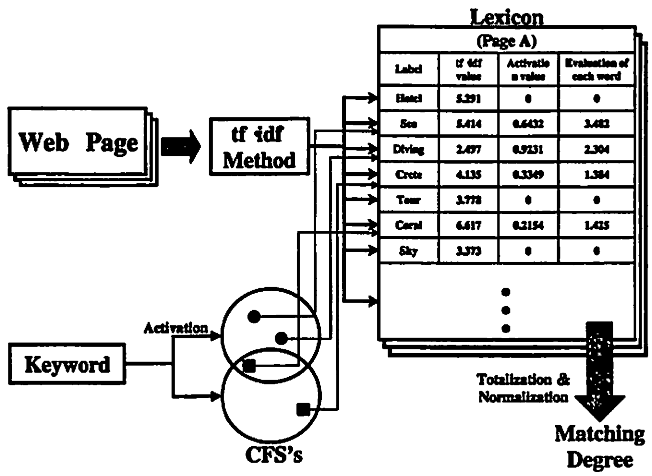


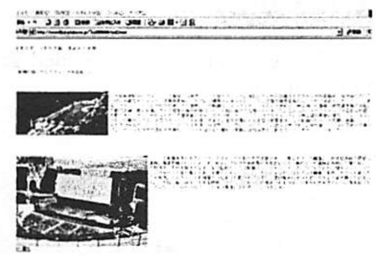
Figure 11. Scheme of the proposed search engine.

Table 2.a Extended keywords.

Ranking	Word	Activation Value
1	The Côte d'Azur	1.0000
2	The Mediterranean Sea	0.9773
3	Famous Resort	0.9187
4	Crete	0.8482
5	Capri	0.6445
6	Anguilla	0.6445
7	Santorini	0.6445
8	Taormina	0.6445
9	Sicily	0.4802
10	Gold Coast	0.0748

Next, abstracts of the retrieved Web pages are listed. Note that, no Web pages were matched by the simple TF-IDF method starting with the keyword input of "famous resort and the Mediterranean Sea (**Figure 12.a**).

Taormina: Ranking 1, Matching degree 1.0



The greatest high-class resort on the island of Sicily. It is located 250 meters above sea level and is known as the "Mediterranean Queen." It has superb views of Mount Etna and the Ionian Sea.

Crete island: Ranking 2, Matching degree 0.83

It is a big island and located in the south. The scenery is different from Mykonos and Santorini islands.

Cote d'Azur: Ranking 3, Matching degree 0.53

Deep-blue coast.

The above results show that our proposed method effectively retrieves information relating to input keywords even when there are no matches with the input keyword itself.

Figure 12.a.

Evaluation 2:

If the CFS unit memorizes knowledge about "vacation" and "sports" such as,

vacation = 1.0/vacance + 0.6/sea + 0.6/sandy beach + 0.6/the South Pacific + ..

sports = 1.0/spots + 0.6/diving + 0.6/trekking + 0.6/golf + ..

then a ranked list of Web pages appears. **Table 2** shows the extended keywords generated by the activation of "vacation" and "sports."

Table 2.b Extended keywords.

Ranking	Word	Activation Value
1	Diving	0.6079
1	Surfing	0.6079
3	Sports	0.5000
3	Vacation	0.5000
5	Golf	0.3039
5	Rock Climbing	0.3039
5	Baseball	0.3039
5	Sea	0.3039
5	Paradise	0.3039
5	Sandy Bearch	0.3039

Next, abstracts of the retrieved Web pages are listed. In contrast, no Web pages were matched by the TF-IDF method using "vacation and sports" (**Figure 12.b**).

Tahiti island: Ranking 1, Matching degree 1.00

Fun for jet skiing and surfing. Diving is also enjoyable.

Boracay island: Ranking 2, Matching degree 0.91

Diving boats and cruising boats come and go. Vacationers have to climb on the boat from the water because there are no piers. It exemplifies resort life.

Rangiroa island: Ranking 3, Matching degree 0.84

Genuine diving! Even snorkeling and a glass-bottomed boat allow glimpses of its mystery.

Figure 12.b.

From the results, we demonstrate the effectiveness of our proposed method. Unlike the first case, the Web pages were not retrieved by place names, but by the activities corresponding to the context of "vacation sports." Jet skiing and surfing were suggested by the CFS as a relevant sports, but baseball was not.

We show that pertinent Web pages can be retrieved independently of the key ward, because even though the region "sport" can encompass a huge number of different activities.

2.1.5. General Observations and Summaries

First, we showed the necessity and also the problems of applying fuzzy sets to information retrieval. Next, we introduced using conceptual fuzzy sets in overcoming those problems, and proposed the realization of conceptual fuzzy sets using Hopfield Networks. Based on above, we proposed the architecture of the search engine which can execute conceptual matching dealing with context-dependent word ambiguity. Finally, we evaluated our proposed method through two simulations of retrieving actual web pages, and compare the proposed method with the ordinary TF-IDF method. We showed that our method could correlate seemingly unrelated input keywords and produce matching Web pages, whereas the simple TF-IDF method could not.

3. Exposure of Illegal Web Sites Using : Conceptual Fuzzy Sets-Based Information Filtering System

Currently, about 1,600 million or more web pages exist on the Internet. People can obtain necessary information from this huge network quickly and easily. On the other hand, various problems have arisen. For example, there are the adult sites, the criminal sites which illegally distribute software (Warez) and music (MP3) and the criminal promotion sites which promote illegal behavior such as making a bomb etc. Therefore, one of the technologies needed currently is the filtering of web sites.

Some software are put to practical use as an internet filter. However, typical software simply match the web sites with illegal URL lists. This approach does not take their contents into consideration at all. These methods lack updating capabilities due to the drastic increase of illegal sites. Additionally, some software eliminates web sites that contain illegal words. They eliminate web sites by calculating confidence of the illegal words in documents by the TF-IDF method. The content-based approach using the TF-IDF method may eliminate any sites that contain harmful words. For example, news sites, which contain harmful words, may be eliminated. In this paper, we propose a filtering system that performs semantic analysis of a web document using conceptual fuzzy sets (CFSs). Our approach concerns not only the appearance of words in a document but also the meanings of words to recognize the harmful nature of a document.

We describe the construction of CFSs using Radial Basis Function (RBF)-like networks in section 3.1. Section 3.2 proposes the filtering system that deals with the semantics of words. Section 3.3 compares our proposed method with the TF-IDF method, and shows the result of filtering simulations of actual web pages. We will conclude the paper in section 3.4.

3.1 Construction of CFSs based on RBF networks

In this section, we will present the construction of CFSs based on RBF networks.

3.1.1 Construction of the network

In the CFSs, words may have a synonymous, antonymous, hypernymous and hyponymous relationship to other words. These relationships are too complicated to be represented in a hierarchical structure. Therefore, we use RBF-like networks to generate CFSs. The image of CFSs is shown in **Figure 13.**

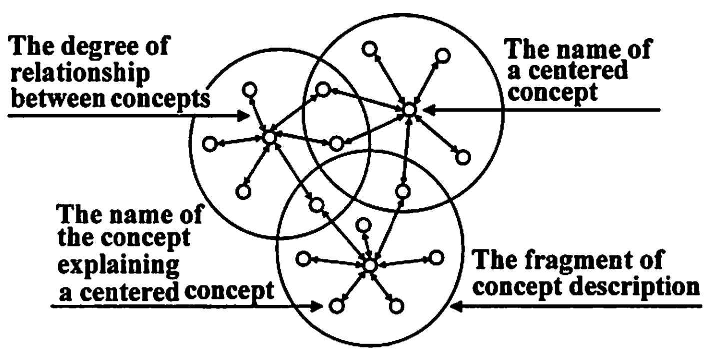


Figure 13. CFSs based on RBF networks

White surrounding concepts explain the centered gray concept. The strength of the links between concepts reflect their degrees of relationship. The centered concept and its connected concepts constitute a fragment of concept description. A CFS is generated by overlapping the fragments of the activated concept desctiption. A CFS expresses the meaning of a concept by the activation values of other concepts in these fragments.

3.1.2 Generation of CFSs

To generate CFSs, concepts are activated using the RBF networks as follows. In general, RBF networks have a structure shown in **Figure 14.**

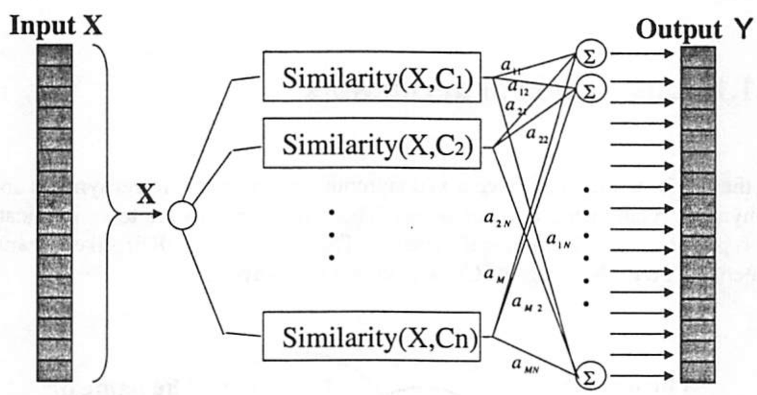


Figure 14. RBF network structure

1) The degree of relationship between a prototype vector c_i (i-th fragment of the concept description) and an input vector x is measured as,

$$\phi(dist(x, c_i))$$

and dist means the distance. Function φ translates the distance to the activation value of the prototype vector. Usually the distance is calculated by Euclidean distance.

$$dist(x, c_i) = \| x - c_i \|.$$

2) The activation values of prototype vectors are weighted with degrees of relationship a_{ij}, and propagate to the relating nodes. So the activation value propagated to j-th node from i-th prototype vector c_i becomes,

$$a_{ij} \times \phi(dist(x, c_i))$$

3) Each node in output layer sums up values translated from all prototype vectors as,

$$\sum_{i=1}^{M} a_{ij} \times \phi(dist(x, c_i))$$

3.2 System description

We developed a system filtering web pages using conceptual fuzzy sets based on RBF networks.

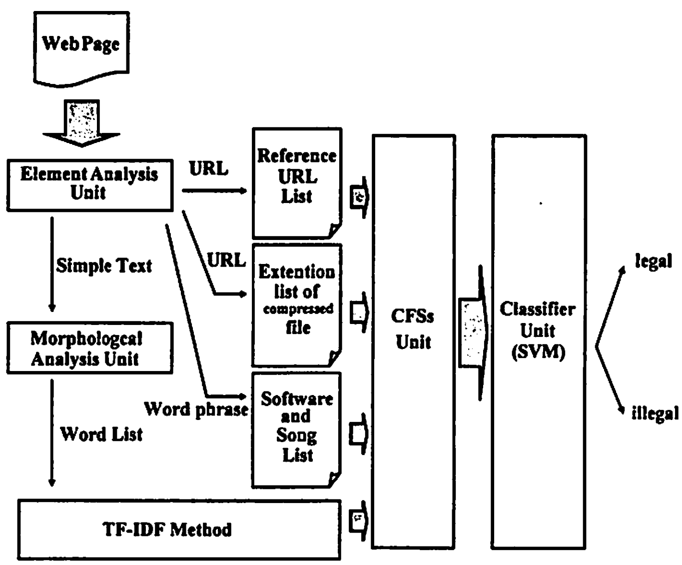


Figure 15. Filtering system using CFSs

3.2.1. Element analysis

In the element analysis unit following two rates and a matching degree are calculated and transferred to the CFSs unit.

- **Rate of links:**

$$\text{Rate of links to compressed files} = \frac{\text{Number of links to compressed files}}{\text{Total number of links}}$$

$$\text{Rate of links to major underground sites} = \frac{\text{Number of links to major illegal sites}}{\text{Total number of links}}$$

We deal with the links to compressed files and the links to major underground sites. It can be conjectured that web pages containing these links have high illegality.

- **Matching degree with name lists:**

The system matches words in the web site with the list of music titles and software names to evaluate tendency toward illegality.

3.2.2 Morphological analysis

The extracted text notes are divided into morphemes. Stop words are excluded. A stop word means a word that occurs frequently despite not having an important meaning and a word that consist of just one character.

3.2.3 Weighting by TF-IDF method

Weights of the words are obtained using general TF-IDF method.

3.2.4 CFSs Unit

A word vector, which consists of TF-IDF values, the link rates and the matching degree, is inputed into the CFSs unit. Propagation of activations occurs from input word vector throughout fragments of the concept descriptions and then abstract concepts "warez", "MP3" and "Emulator" are recognized. CFSs units consists of fragments of the concept descriptions, such as "warez", "MP3", etc as shown in **Figure 16.**

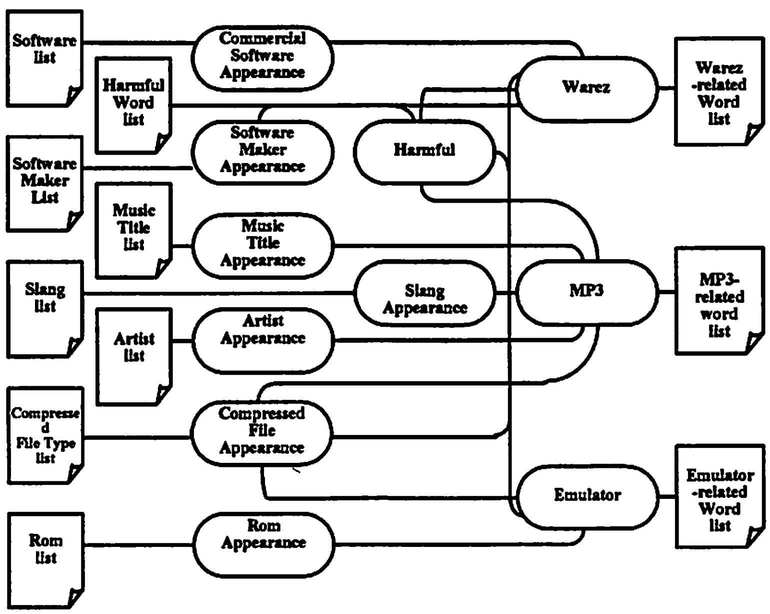


Figure 16. Concept description in the CFS unit

For example, the concept description of "MP3" is defined as shown in **Table 3.**

Table 3. Concept description of "MP3"

Concepts	Relationship
mp3	1
mp3s	0.9
mp3z	0.9
<song title>	0.8
<artist name>	0.8
napster	0.8
winmx	0.8
artist	0.8
song	0.8
music	0.7
album	0.7
single	0.7
title	0.6
cd	0.6
cds	0.6
cdz	0.6
<compressed file>	0.5
zip	0.5
band	0.5

Only highly ranked concepts are displayed here. It should be noticed that the concept description includes some signatures, such as link rates, to reflect human subjectivity.

3.2.5 Classifier Unit

We used the Support Vector Machine (SVM) as a classifier. The classifier unit inputs the activation values of words in the CFSs unit and distinguishes illegal sites from legal ones.

3.3 Evaluations Procedure

In this study, we randomly selected 300 actual web sites as samples for evaluation, and we compared the proposed method with the support vector machine. The samples contained 85 illegal sites. We assumed seven types of the illegal sites shown in **Table 4.** This classification is not based on the law strictly but on common sense. We evaluated the system by filtering Warez, Emulator and MP3 sites in this study.

Table 4. Seven types of illegal sites

Group	A classification criterion
Warez	Illegal distribution and sale of commercial softwares
Emulater	Illegal distribution of softwares, such as consumer games and video games
MP3	Disutribution of music data which infringe on copyright
Adult	Dirty depictions and expressions
Hack & Crack	Distribution of illegal hacking and cracking softwares Instruction of technical know-how
Drug & Gun	Sale of drugs and guns Introduction of acquisition route
Killing	Expressions about murder, violete depiction, etc.

3.3.1 Results

The results of the experiments are shown in **Table 5** and **Table 6.**

Table 5. The classification results

	proposed system		TF-IDF method	
	success	failure	success	failure
illegal document	81	4	74	11
legal document	214	1	215	0
all document	295	5	289	11

Table 6. Comparison by E measure

		300 documents	
		proposed system	TF-IDF method
illegal document	precision	0.9878	1.0000
	recall	0.9529	0.8706
	E measure	0.0299	0.0692
ligal document	precision	0.9817	0.9556
	recall	0.9953	1.0000
	E measure	0.0115	0.0227

Table 6 compares the experimental results from the viewpoint of precision, recall and E measure. The proposed method exceeded simple TF-IDF method in finding illegal sites, although it is inferior in the case of legal sites.

Here is an example (in **Figure 17**) which is distinguished correctly by the proposed system but not by simple TF-IDF method.

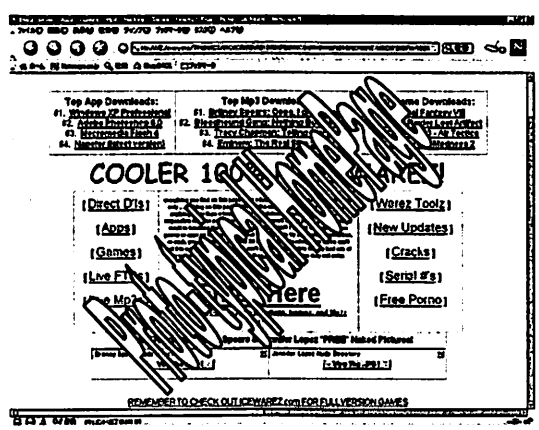


Figure 17. Page evaluated by the proposed system.

Highly ranked 20 words of TF-IDF value are shown in **Table 7**.

Table 7. Word vector of Web site

Ranking	Word	TF- IDF value
1	1spc	0.10751
2	crusader	0.10146
3	rate	0.09140
4	jpg	0.05966
5	flt	0.05759
6	com1	0.05553
7	apps	0.04492
8	port	0.03896
9	tag	0.03660
10	jennifer	0.02922
11	britney	0.02885
12	pub	0.02820
13	nov	0.02722
14	razor1911	0.02687
15	3spcs	0.02687
16	echelon	0.02687
17	wolfenstein	0.02598
18	systran	0.02564
19	simak	0.02304
20	crackin	0.02304

The reason the TF-IDF method failed could be due to the fact that this Web page is very large. It also contains a legal part in the latter half, and thus the TF-IDF values indicating illegality were decreased. Another reason is interpretation of link rates. **Table 8** indicates highly ranked 10 words used as the input of CFSs Unit.

Table 8. Input values to CFSs

Ranking	Word	Input-value (TF-IDF value)
1	<software list>	0.13333
2	apps	0.04492
3	iso	0.02192
4	warez	0.01934
5	psx	0.01568
6	mp3	0.01424
7	game	0.01280
8	ftp	0.01276
9	album	0.01074
10	get	0.01052

Table 9. Output values of CFSs

Ranking	Word	Activation value
1	<software list>	0.14441
2	apps	0.05646
3	warez	0.03673
4	iso	0.02884
5	game	0.02341
6	gamez	0.02130
7	mp3	0.02020
8	psx	0.01845
9	appz	0.01798
10	serialz	0.01787

Because the famous software names appeared frequently, activation values of the concepts besides "warez", such as gamez, appz and serialz, arose. (**Table 9**). It is considered that CFSs can recognize that the concept "warez" fusing word frequency and link information reflecting human subjectivity.

3.4 General Observations and Summaries

In this section, we applied the CFSs using RBF networks. Moreover, we proposed a system which is capable of filtering harmful web sites. We showed that the semantic interpretation of the concept by CFSs exceeded the TF-IDF method which is based on the superficial statistical information.

However, the proposed system has been evaluated using the limited number of target web documents. In our future work, we need to strengthen the conceptual descriptions and generalizations of CFSs that can be used in the entire Internet.

3.6 Fuzzy-TF.IDF

The use of Fuzzy-tf-idf is an alternative to the use of the conventional tf-idf. In this case, the original tf-idf weighting values will be replaced by a fuzzy set rather than original crisp value. To reconstruct such value both ontology and similarity measure can be used. To develop ontology and similarity one can used the conventional Latent Semantic Indexing (LSI) or Fuzzy-LSI (Nikravesh and Azvine 2002). The fuzzy-LSI (**Figure 18**), fuzzy-TF-IDF, and CFS can be used through an integrated system to develop fuzzy conceptual model for intelligent 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*** (**Figure19**).

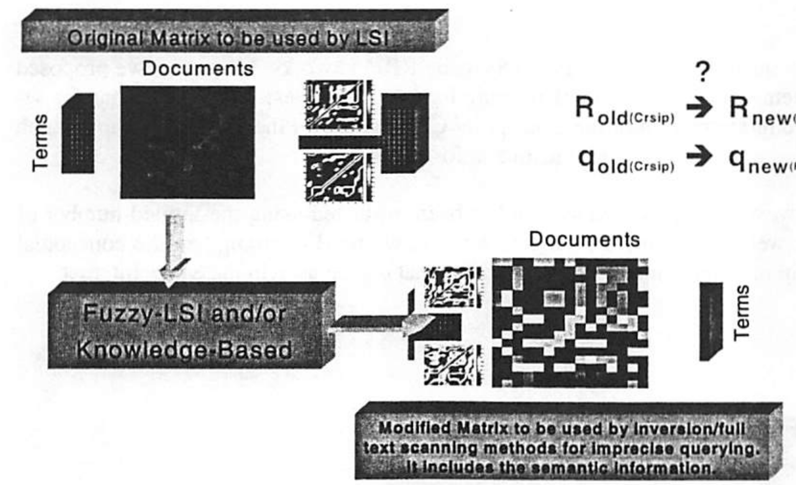


Figure 18. Fuzzy-Latent Semantic Indexing-Based Conceptual Technique

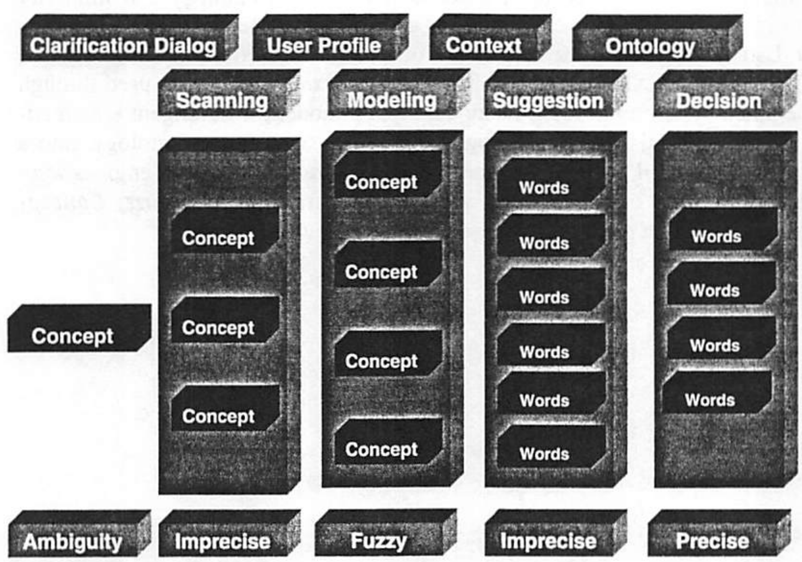


Figure 19. Fuzzy Conceptual Matching and Human Mental Model

4. Conceptual Fuzzy Sets-Based Navigation System for Yahoo!

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 (Zadeh 2001b and 1999).

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. Already explosive amount of users on the Internet is estimated over hundreds of millions. For example over 30 million new users visiting Google™ every month. 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.

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 (Altavista, Excite, HotBot, Infoseek, and Lycos), 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 (For all five search engines; Mean Hit: %15, Proximity: %21.4, and Location: %50.2). The effectiveness of the classification is defined based on the precision and recall. 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. 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 or lower value. 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.

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. While searches may retrieve thousands of hits, finding relevant partial matches and query relevant information with deductive capabilities might be a problem. 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 Fuzzy-LSI for automatic information retrieval and search with partial matches are required.

4.1 Navigation System for Yahoo!

Many search engines such as *Yahoo!* classify a large number of web sites into their own large hierarchical categories (directories). Although category menus are provided for users, the users don't commonly know the hierarchical structure nor do they understand which item (categories) on the menus to select to find documents they want.

In this section, we propose a navigation system which conceptually matches input keywords with all paths from a root category to leaf categories. Input keywords don't always match words on category menus directly. The proposed system conceptually matches keywords with paths by taking the meaning of a path into consideration and by expanding keywords. For conceptual matching, we use CFSs based on RBF networks.

We describe the CFSs based on RBF networks in section 4.2 and our navigation system in section 4.3. In section 4.4, we present our experiments and results.

4.2 Conceptual fuzzy sets

For conceptual matching, we use conceptual fuzzy sets (CFSs) based on radial basis function (RBF) networks. In a CFS, the meaning of a concept is represented by the distribution of the activation values of the other concepts. To expand input keywords, the propagation of activation values is carried out recursively.

4.2.1 RBF-based CFSs

In the CFSs, relationships between words may be synonymous, antonymous, hypernymous and hyponymous . These relationships are too complicated to be represented in a hierarchical structure. Therefore, we use RBF-like networks to generate CFSs. **Figure 4** shows the image of CFSs based on RBF networks.
Each node corresponds to one concept and input/output is represented as the activation values of nodes.

RBF networks originally learn the prototype vectors and the weights between nodes from data. In this section, however, we generate CFSs using a concept base which we made manually.

4.2.2 Concept base

A concept base is a knowledge base which stores words represented in other words with their degrees of relationship. Although word co-occurrence measures are widely used to calculate the degrees of relationship between words, we use an original method based on some rules found in Japanese language dictionaries.

The rules are:

1. A word that has highly relative meaning to a headword usually appears first.

2. If a single word represents the meaning of a headword, it is a synonym or it has strong relationship to the headword.

We add some words that are relative to headwords because the networks are too sparse if we construct them only with a Japanese language dictionary. Then the degrees of relationships between words are calculated with the rules above. **Table 10** shows some examples of headwords and their relative words with the degrees of relationships.

Table 10. Example of degrees of relationship

headword	magazine		book	
words explaining the headword	magazine	1.0	book	1.0
	book	0.8	publication	0.84
	bookstore	0.72	magazine	0.72
	publication	0.6	journal	0.7
	newspaper	0.5	bookstore	0.64
	information	0.4	dictionary	0.6

We generate the CFSs considering each vector of a headword as a prototype vector of RBF networks, and each degree of relationship between words as a weight between output and corresponding unit.

4.3 Navigation system

In this section, we describe our navigation system which conceptually matches input keywords and all paths. The system consists of a CFS unit, a path base, a matching unit and a user interface. **Figure 20** shows the architecture of the system.

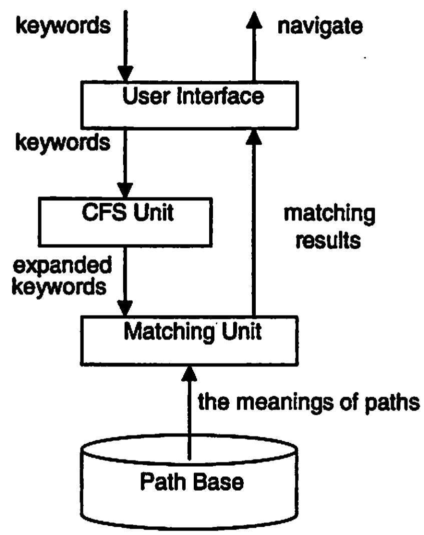


Figure 20. Navigation System Architecture

4.3.1 CFS unit

When keywords are inputed into the CFS unit, propagation of activation values occurs from the keywords. It results in the distribution of the activation values of the other words and represents the concept of the keywords. The propagation is carried out recursively several times to associate with relative concepts. This recursive propagation enables the association of concepts which are connected indirectly with input keywords (**Figure 21**).

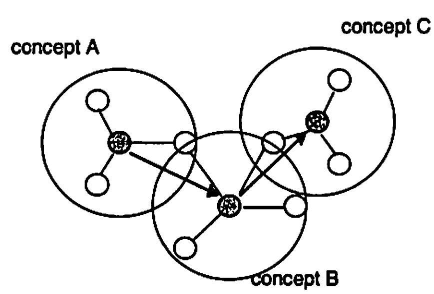


Figure 3. Image of association indirectly connected concepts

4.3.2. Path base

A path is a sequence of category labels from a root category to a leaf category. We take the meaning of a path into consideration to search paths to appropriate categories.

The meaning of a path is the result of expansion in the CFS unit from the category labels in the path. The path base stores all the paths and their corresponding meanings. Figure 22 shows the image of expanded paths.

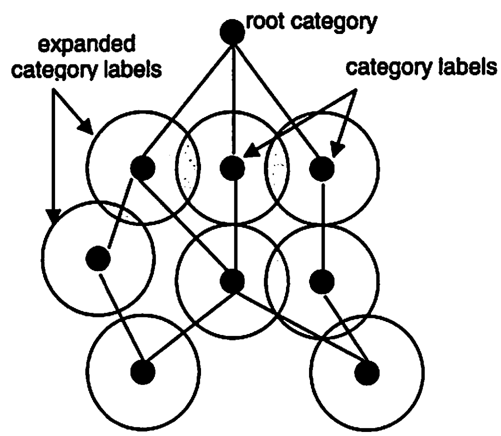


Figure 4. Image of expanded

4.3.3 Matching unit

The matching unit calculates the similarity between expanded input keywords and each expanded path. We

use the cosine measure to calculate the similarity:

$$Sim(C_k, C_{p_i}) = \sum_{j=0}^{N} w_{kj} \times w_{p_i j}$$

where $C_k = (w_{k1}, w_{k2}, \ldots, w_{kN})$ is expanded keywords, and $C_{p_i} = (w_{p_i 1}, w_{p_i 2}, \ldots, w_{p_i N})$ is the i th expanded path.

4.3.4 User interface

The user interface displays recommended paths according to matching degrees to navigate the user.

4.4 Experiments

In this section, we evaluate the effectiveness of our navigation system using test data shown in **Table 11. Table 12** shows some examples of paths. They referred *Yahoo! JAPAN.*

First, we evaluated how many times the propagation of activation values should be carried out in the CFS unit. Second, we actually search the paths and evaluate the results.

Table 11. Test data

the number of paths	213
the number of words	803
the number of headwords	214

Table 3. Example of paths

Business and Economy > Cooperatives > Architecture > Software
Computer and Internet > Software > Internet
Entertainment > Music > Composition
Government > Law > Tax
Science > Physics > Libraries

4.4.1 Determine the repeat number of propagation

Figure 23 shows the changes of activation values of some words with "personal computer" and "book" as input to the CFS unit.

The activation value of the word "magazine" gets higher as the propagation is carried out and is at the peak in the third to fifth iteration. The word "magazine", which highly relates to "book", is associated by iterating propagation of activation values in CFS unit. The activation value of the word "information" is also at the peak in the third to fifth iteration. Although the word "information" is not connected directly to "personal computer" nor "book", the iteration of the propagation of activation values enables the association of the word.

4.4.2 Search the paths

We assume that a user is interested in books about personal computers, and then he/she inputs "personal computer" and "book" as keywords. We fixed the repeat number of propagation in the CFS unit on three times and searched the paths with these keywords. The result is shown in **Table 13.**

The top ranked path leads to the category which is assigned to web sites of online computer bookstores. The system could search the path that seems to be the best for the input keywords.

Table 13. Matching results

Business and Economy > Cooperatives > Books > Booksellers > Computers Similarity = 0.951106
Business and Economy > Cooperatives > Books > Booksellers > Magazines Similarity = 0.945896
Business and Economy > Cooperatives > Books > Booksellers > Movies Similarity = 0.918033
Business and Economy > Cooperatives > Books > Booksellers > Architecture Similarity = 0.9093
Business and Economy > Cooperatives > Books > Booksellers > Literature Similarity = 0.904156

Note that the first item in the best path is "Business and Economy", which may be unexpected for him/her to have to click on to reach the computer bookstores. Our system could recommend such a path that lets the user find categories he/she wants.

However, all the top five paths in the search result lead to categories about books. The reason of this may be that the concept base includes too many words about books.

4.4.3 3D user interface

We have developed a 3D user interface to navigate users effectively. The interface displays hierarchical structure in the same manner as Cone Tree (Robertson et al 90 and 91). **Figure 24** is a screenshot of the user interface. Users can easily understand their position in large categorical hierarchy and the system can prevent them from getting lost. And for users who want to get more detail, functions such as rotation and zoom are also provided.

Paths with high similarities to input keywords are highlighted and the system can help users to reach appropriate categories.

4.5 General Observations and Summaries

We proposed a navigation system which conceptually matches input keywords and paths using CFSs based on RBF networks. Taking the meaning of a path into consideration and propagating activations of concepts recursively in CFS unit to associate relative words with input keywords enabled the system to search the path leading to an appropriate category.

However, the following are some problems which require further study:

- The scale of the system is small.
- The associations in CFS unit are affected by un-uniformity of the concept base.
- The number of propagation of activation values in CFS unit is empirical.

In this study, we used the cosine measure to calculate the similarity. In the future work, we intent to use other similarity measures (as shown in **Table 14** and **Table 15**) especially perception-based and fuzzy-based similarity measures (Nikravesh 2002, Nikravesh et al. 2002).

Table 14. Five commonly used measure of similarity and association in IR

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

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

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

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

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

Disimilarity Coefficeint: $\frac{|X \Delta Y|}{|X|+|Y|} = 1 - \textit{Dice's}$ Coefficient

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

Table 15. Term-Document Matrix Representation (R) and Similarity Measure (For definition of the terms used in this table refer to Ref. (Zobel and Moffat).

$$\textit{Similarity - measure}: \mathrm{S}_{q,d} = f(w_{q,t}, w_{d,t}, W_q, W_d, \tau_{q,d}, C, w_t, r_{d,t})$$

$$\textit{Term - weights}: w_t = f(N, f_t, f^m, f_{d,t}, F_t, n_t, s_t)$$

$$\textit{Document - term - weights}: w_{d,t} = f(r_{d,t}, w_t)$$

$$\mathrm{Re}\textit{lative - term - frequencies}: r_{d,t} = f(t, \tau_d, f_{d,t}, f_d^m, W_d)$$

$$\textit{Document - lengths and query - lengths}: W_d = f(t, \tau_d, w_{d,t}, f_d, s, W_d')$$

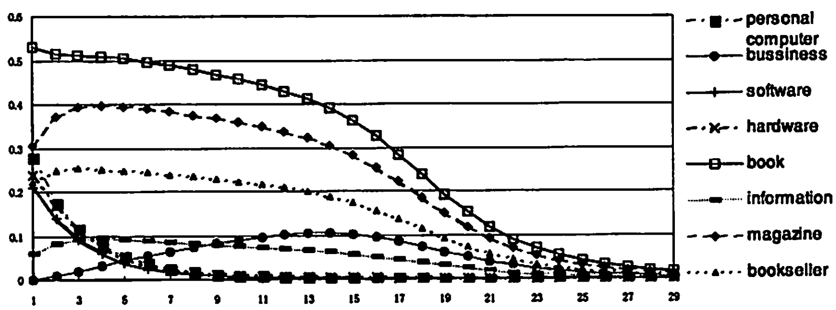


Figure 23. Changes of activation values

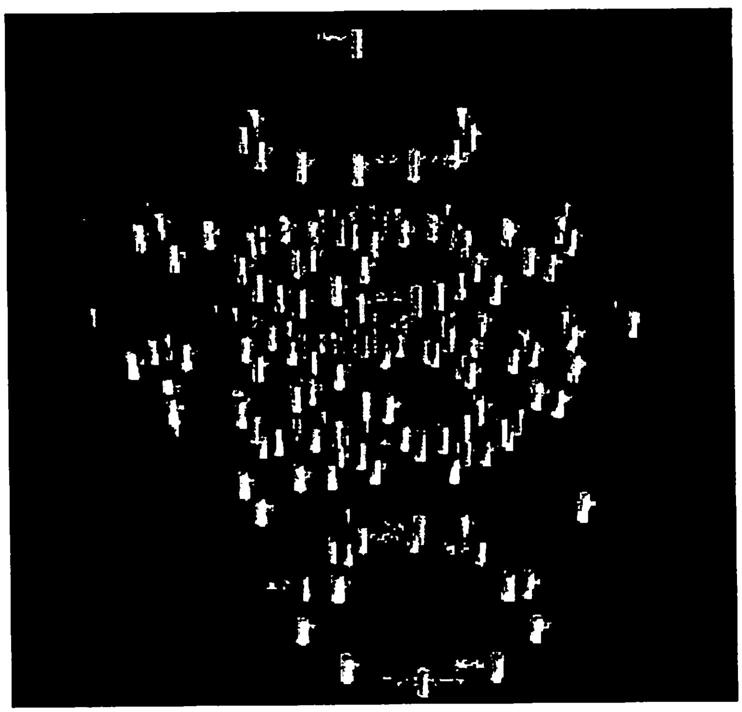


Figure 24. 3D user interface

5 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. Following 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:

- *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
- *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 SpaceDeductive Capabilities
    - Imprecise-Querying
    - *Ambiguity Resolution through Clarification Dialog*
    - *Precisiated Natural Languages (PNL)*

- ***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 QueryContextual Categorization

- ***Fuzzy Logic and Internet; Fundamental Research:***
  - Computing with Words (CW)
  - Computational Theory of Perception (CTP)
  - Precisiated Natural Languages (PNL)

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

- ***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

- ***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.

6 Conclusions

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.

In this work, we proposed a search engine which conceptually matches input keywords and web pages. The conceptual matching is realized by context-dependent keyword expansion using conceptual fuzzy sets. First, we show the necessity and also the problems of applying fuzzy sets to information retrieval. Next, we introduce the usefulness of conceptual fuzzy sets in overcoming those problems, and propose the realization of conceptual fuzzy sets using Hopfield Networks. We also propose the architecture of the search engine which can execute conceptual matching dealing with context-dependent word ambiguity. Finally, we evaluate our proposed method through two simulations of retrieving actual web pages,and compare the proposed method with the ordinary TF-IDF method. We show that our method can correlate seemingly unrelated input keywords and produce matching Web pages, whereas the TF-IDF method cannot.

Currently on the Internet there exists a host of illegal web sites which specialize in the distribution of commercial software and music. This section proposes a method to distinguish illegal web sites from legal ones not only by using tf-idf values but also to recognize the purpose/meaning of the web sites. It is achieved by describing what are considered to be illegal sites and by judging whether the objective web sites match the description of illegality. Conceptual fuzzy sets (CFSs) are used to describe the concept of illegal web sites. First, we introduced the usefulness of CFSs in overcoming those problems, and propose the realization of CFSs using RBF-like networks. In a CFS, the meaning of a concept is represented by the distribution of the activation values of the other nodes. Because the distribution changes depend on which labels are activated as a result of the conditions, the activations show a context-dependent meaning. Next, we proposed the architecture of the filtering system. Additionally, we compared the proposed method with the tf-idf method with the support vector machine. The e-measures as a total evaluation indicate that the proposed system showed better results as compared to the tf-idf method with the support vector machine.

Finaly, we proposed a menu navigation system which conceptually matches input keywords and paths. For conceptual matching, we used conceptual fuzzy sets (CFSs) based on radial basis function (RBF) networks. In a CFS, the meaning of a concept is represented by the distribution of the activation values of the other concepts. To expand input keywords, the propagation of activation values is carried out recursively. The proposed system recommends users paths to appropriate categories. We used 3D user interface to navigate users.

7. Future Works

7.1 TIKManD (Tool for Intelligent Knowledge Management and Discovery)

In the future work, we intent to develop and deploy an intelligent computer system is called *"TIKManD (Tool for Intelligent Knowledge Management and Discovery)"*.

The system can mine Internet homepages, Emails, Chat Lines, and/or authorized wire tapping information (which may include Multi-Lingual information) to recognize, conceptually match, and rank potential terrorist and criminal activities (both common and unusual) by the type and seriousness of the activities. This will be done automatically or semi-automatically based on predefined linguistic formulations and rules defined by experts or based on a set of known terrorist activities given the information provided through law enforcement databases (text and voices) and huge number of "tips" received immediately after the attack. Conceptual Fuzzy Set (CFS) model will be used for intelligent information and knowledge retrieval through conceptual matching of text, images and voice (here defined as "Concept"). The CFS can be also used for constructing fuzzy ontology or terms relating the context of the investigation (Terrorism or other criminal activities) to resolve the ambiguity. This model can be used to calculate conceptually the degree of match to the object or query. In addition, the ranking can be used for intelligently allocating resources given the degree of match between objectives and resources available (Nikravesh et al. 2002, Nikravesh 2002).

The use of the Conceptual Fuzzy Set (CFS) is a necessity, given the ambiguity and imprecision of the "concept" in law enforcement databases and information related to terrorism, which may be described by Multi-Lingual textual, images and voice information. In the CFS approach, the "concept" is defined by a series of keywords with different weights depending on the importance of each keyword. Ambiguity in concepts can be defined by a set of imprecise concepts. Each imprecise concept in fact can be defined by a set of fuzzy concepts. The fuzzy concepts can then be related to a set of imprecise words given the context. Imprecise words can then be translated into precise words given the ontology and ambiguity resolution through clarification dialog. By constructing the ontology and fine-tuning the strength of links (weights), we could construct a fuzzy set to integrate piecewise the imprecise concepts and precise words to define the ambiguous concept.

7.2 Web Intelligence: Google™ and Yahoo! Concept-Based Search Engine

There are two type of search engine that we are interested and are dominating the Internet. First, the most popular search engines that are mainly for unstructured data such as Google ™ and Teoma which are based on the concept of Authorities and Hubs. Second, search engines that are task spcifics such as 1) Yahoo!: manually-pre-classified, 2) NorthernLight: Classification, 3) Vivisimo: Clustering, 4) Self-organizing Map: Clustering + Visualization and 5) AskJeeves: Natural Languages-Based Search; Human Expert.

Google uses the PageRank and Teoma uses HITS (Ding et al. 2001) for the Ranking. **Figure 25** shows the Authorities and Hubs concept and the possibility of comparing two homepages.

Figures 26 shows the possible model for similarity analysis is called "fuzzy Conceptual Similarity". **Figure 27** shows the matrix representation of Fuzzy Conceptual Similarity model. **Figure 28** shows the evolution of the Term-Document matrix. **Figure 29** shows the structure of the Concept-based Google ™ search engine for Multi-Media Retrieval . Finally, **Figure 30** shows the structure of the Concept-Based Intelligent Decision Analysis. To develop such models, state-of-the-art computational intelligence techniques are needed. These include and are not limited to:

- Latent-Semantic Indexing and SVD for preprocessing,
- Radial-Basis Function Network to develop concepts,
- Support Vector Machine (SVM) for supervised classification,
- fuzzy/neuro-fuzzy clustering for unsupervised classification based on both conventional learning techniques and Genetic and Reinforcement learning,
- non-linear aggregation operators for data/text fusion,
- automatic recognition using fuzzy measures and a fuzzy integral approach
- self organization map and graph theory for building community and clusters,
- both genetic algorithm and reinforcement learning to learn the preferences,

- fuzzy-integration-based aggregation technique and hybrid fuzzy logic-genetic algorithm for decision analysis, resource allocation, multi-criteria decision-making and multi-attribute optimization.

- text analysis: next generation of the Text, Image Retrieval and concept recognition based on soft computing technique and in particular Conceptual Search Model (CSM). This includes

    - Understanding textual content by retrieval of relevant texts or paragraphs using CSM followed by clustering analysis.
    - Hierarchical model for CSM
    - Integration of Text and Images based on CSM
    - CSM Scalability, and
    - The use of CSM for development of

        - Ontology
        - Query Refinement and Ambiguity Resolution
        - Clarification Dialog
        - Personalization-User Profiling

Acknowledgement

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

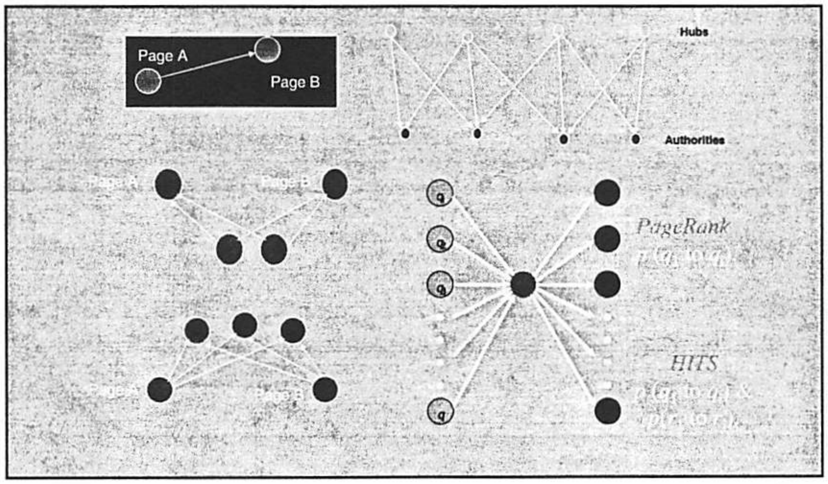


Figure 25. Similarity of web pages.

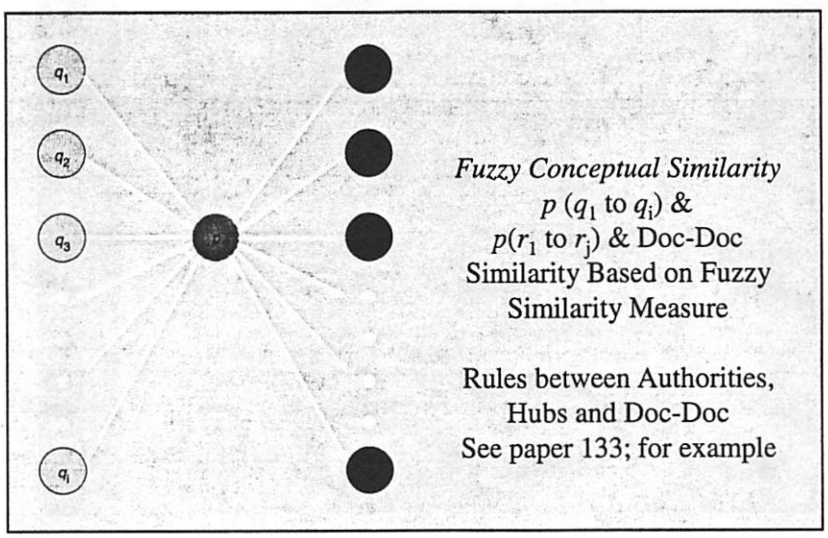


Figure 26. Fuzzy Conceptual Similarity

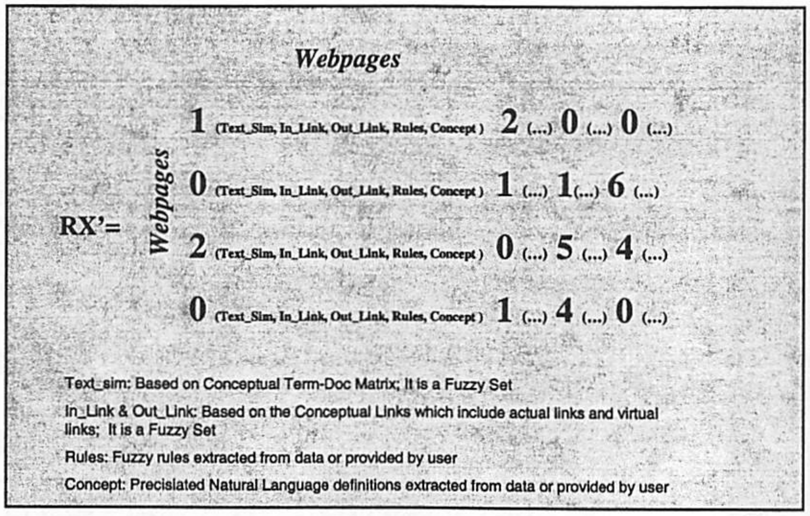


Figure 27. Matrix representation of Fuzzy Conceptual Similarity model

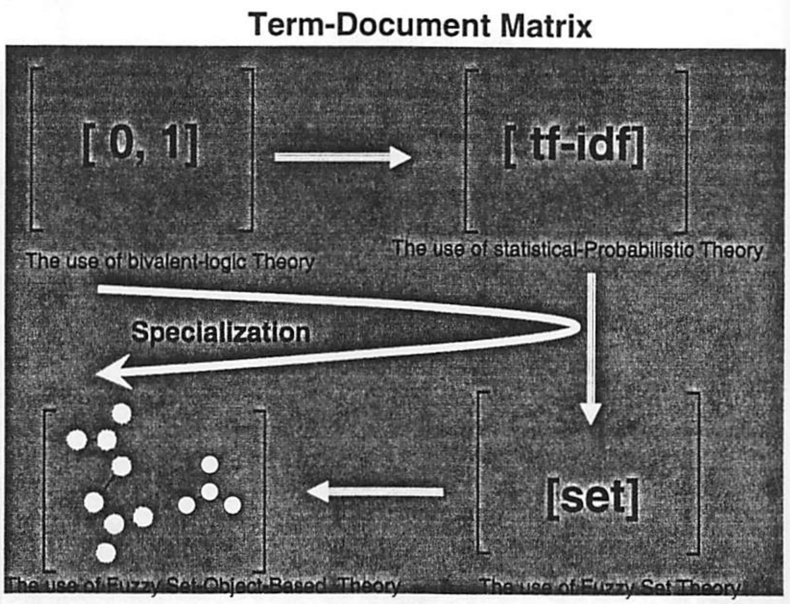


Figure 28. Evolution of Term-Document Matrix representation

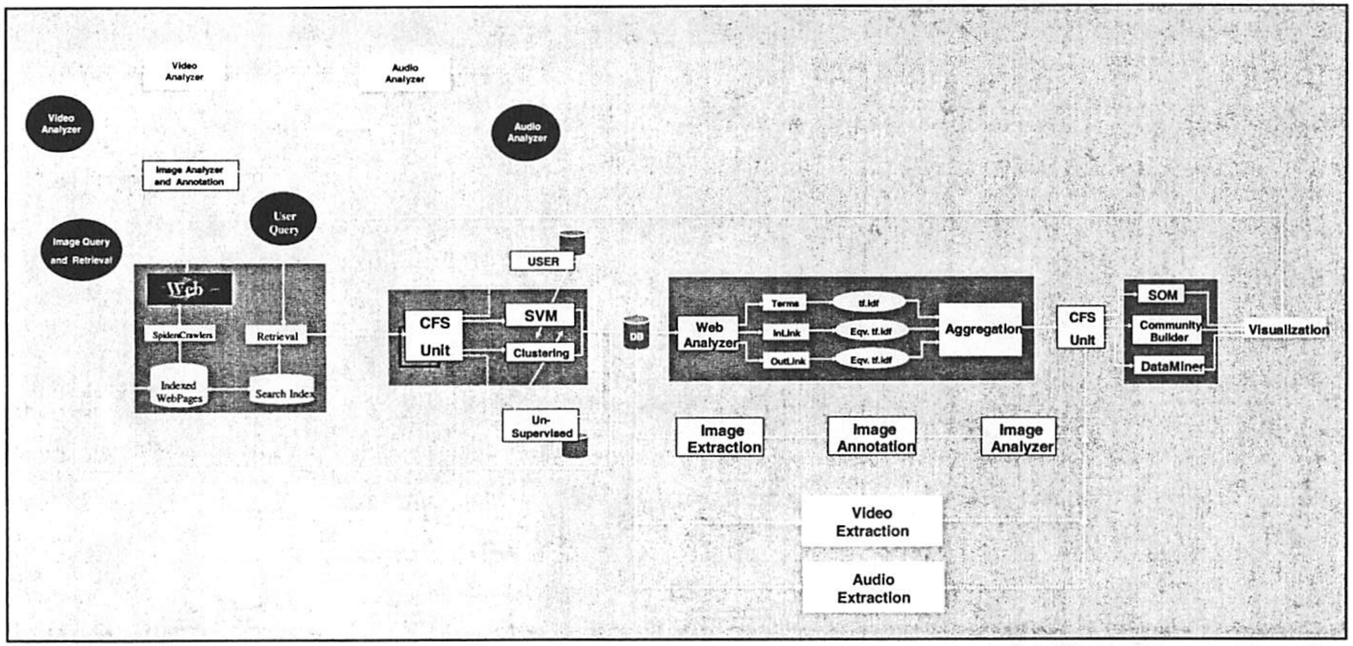


Figure 29. Concept-Based Google™ Search Engine for Multi-Media Retrieval

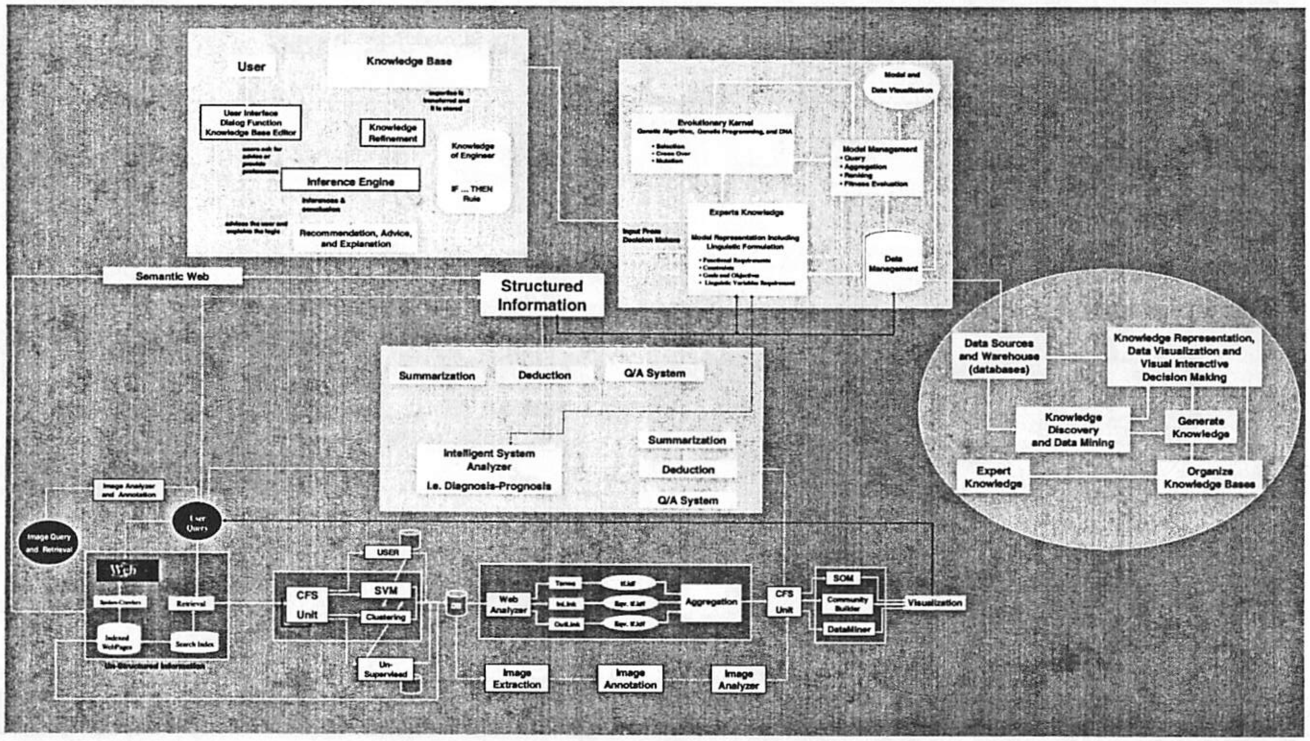


Figure 30. concept-Based Intelligent Decision Analysis

References

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.

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.

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.

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

Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha and Horst D. Simon, PageRank, HITS and a Unified Framework for Link Analysis. LBNL Tech Report 50007. Nov 2001. Proc. of 25th ACM SIGIR Conf. pp.353 354, 2002 (poster), Tampere, Finland

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.

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.

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).

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. 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.

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.

M. Nikravesh, V. Loia,, and B. Azvine, Fuzzy logic and the Internet (FLINT), Internet, World Wide Web, and Search Engines, to be appeared in International Journal of Soft Computing-Special Issue in fuzzy logic and the Internet , 2002

M. Nikravesh, Fuzzy Conceptual-Based Search Engine using Conceptual Semantic Indexing, NAFIPS-FLINT 2002, June 27-29, New Orleans, LA, USA

M. Nikravesh and B. Azvin, Fuzzy Queries, Search, and Decision Support System, to be appeared in International Journal of Soft Computing-Special Issue in fuzzy logic and the Internet , 2002

M. Nikravesh, V. Loia,, and B. Azvine, Fuzzy logic and the Internet (FLINT), Internet, World Wide Web, and Search Engines, to be appeared in International Journal of Soft Computing-Special Issue in fuzzy logic and the Internet , 2002

M. Nikravesh, Fuzzy Conceptual-Based Search Engine using Conceptual Semantic Indexing, NAFIPS-FLINT 2002, June 27-29, New Orleans, LA, USA

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.

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.

George G. Robertson, Stuart K. Card, and Jock D. Mackinlay, "Information Visualization Using 3D Interactive Animation", *Communications of the ACM*, Vol.36 No.4, pp.57-71, 1990.

George G. Robertson, Jock D. Machinlay, and Stuart K. Card, "Cone Trees: Animated 3D Visualizations of Hierarchical Information", *Proceedings of CHI '91*, pp.189-194.

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).

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.

John 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 2001b, 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.

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.

J.Zobel and A. Moffat, Exploring the Similarity Space, http://www.cs.mu.oz.au/~alistair/exploring/.

Enhancing the Power of the Internet

1. From Search Engines to Question-Answering Systems: The Need for New Tools (L.A. Zadeh)

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."

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.

2. Web Intelligence: Conceptual Search Engine and Navigation (M. Nikravesh)

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:

i· 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.

ii· Another, and less obvious, is deduction in an unstructured and imprecise environment given the huge stream of complex information.

One can use clarification dialog, user profile, context, and ontology, into an integrated frame work to design a more intelligent search engine. The model will be used for intelligent information and knowledge retrieval through conceptual matching of text. The selected query doesn't need to match the decision criteria exactly, which gives the system a more human-like behavior. The model can also be used for constructing ontology or terms related to the context of search or query to resolve the ambiguity. 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.

It is also possible to automate ontology generation and document indexing using the terms similarity based on Conceptual-Latent Semantic Indexing Technique (CLSI). 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. It is also possible to generate conceptual documents similarity map that can be used for intelligent search engine based on CLSI, 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.

Given the ambiguity and imprecision of the "concept" in the Internet, which may be described by both textual and image information, the use of Fuzzy Conceptual Matching (FCM) is a necessity for search engines. In the FCM approach, the "concept" is defined by a series of keywords with different weights depending on the importance of each keyword. Ambiguity in concepts can be defined by a set of imprecise concepts. Each imprecise concept in fact can be defined by a set of fuzzy concepts. The fuzzy concepts can then be related to a set of imprecise words given the context. Imprecise words can then be translated into precise words given the ontology and ambiguity resolution through clarification dialog. By constructing the ontology and fine-tuning the strength of links (weights), we could construct a fuzzy set to integrate piecewise the imprecise concepts and precise words to define the ambiguous concept.

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.

2.1 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. Following 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:

- *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
- *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 SpaceDeductive Capabilities

    - Imprecise-Querying
    - *Ambiguity Resolution through Clarification Dialog*
    - *Precisiated Natural Languages (PNL)*
- ***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 QueryContextual Categorization
- ***Fuzzy Logic and Internet; Fundamental Research:***
    - Computing with Words (CW)
    - Computational Theory of Perception (CTP)
    - Precisiated Natural Languages (PNL)

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

- ***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
- ***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.

2.2 Conclusions

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 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.

2.3 Future Works

2.3.1 TIKManD (Tool for Intelligent Knowledge Management and Discovery)

In the future work, we intent to develop and deploy an intelligent computer system is called ***"TIKManD (Tool for Intelligent Knowledge Management and Discovery)"***.

The system can mine Internet homepages, Emails, Chat Lines, and/or authorized wire tapping information (which may include Multi-Lingual information) to recognize, conceptually match, and rank potential terrorist and criminal activities (both common and unusual) by the type and seriousness of the activities. This will be done automatically or semi-automatically based on predefined linguistic formulations and rules defined by experts or based on a set of known terrorist activities given the information provided through law enforcement databases (text and voices) and huge number of "tips" received immediately after the attack. Conceptual Fuzzy Set (CFS) model will be used for intelligent information and knowledge retrieval through conceptual matching of text, images and voice (here defined as "Concept"). The CFS can be also used for constructing fuzzy ontology or terms relating the context of the investigation (Terrorism or other criminal activities) to resolve the ambiguity. This model can be used to calculate conceptually the degree of match to the object or query. In addition, the ranking can be used for intelligently allocating resources given the degree of match between objectives and resources available.

2.3.2 Google™ and Yahoo! Concept-Based Search Engine

There are two type of search engine that we are interested and are dominating the Internet. First, the most popular search engines that are mainly for unstructured data such as Google ™ and Teoma which are based on the concept of Authorities and Hubs. Second, search engines that are task spcifics such as 1) Yahoo!: manually-pre-classified, 2) NorthernLight: Classification, 3) Vivisimo: Clustering, 4) Self-organizing Map: Clustering + Visualization and 5) AskJeeves: Natural Languages-Based Search; Human Expert.

Google uses the PageRank and Teoma uses HITS (Ding et al. 2001) for the Ranking. To develop such models, state-of-the-art computational intelligence techniques are needed. These include and are not limited to:

- Latent-Semantic Indexing and SVD for preprocessing,
- Radial-Basis Function Network to develop concepts,
- Support Vector Machine (SVM) for supervised classification,
- fuzzy/neuro-fuzzy clustering for unsupervised classification based on both conventional learning techniques and Genetic and Reinforcement learning,
- non-linear aggregation operators for data/text fusion,
- automatic recognition using fuzzy measures and a fuzzy integral approach
- self organization map and graph theory for building community and clusters,
- both genetic algorithm and reinforcement learning to learn the preferences,
- fuzzy-integration-based aggregation technique and hybrid fuzzy logic-genetic algorithm for decision analysis, resource allocation, multi-criteria decision-making and multi-attribute optimization.
- text analysis: next generation of the Text, Image Retrieval and concept recognition based on soft computing technique and in particular Conceptual Search Model (CSM). This includes
    - Understanding textual content by retrieval of relevant texts or paragraphs using CSM followed by clustering analysis.
    - Hierarchical model for CSM
    - Integration of Text and Images based on CSM
    - CSM Scalability, and
    - The use of CSM for development of
        - Ontology
        - Query Refinement and Ambiguity Resolution
        - Clarification Dialog
        - Personalization-User Profiling

3. Conceptual Fuzzy Sets and its Application to Web Information Retrieval (T. Takagi)

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. We applied the conceptual fuzzy set to web information retrieval based on its capability to measure conceptual distance between documents.

3.1 Conceptual Fuzzy Sets

Main cause of vagueness is ambiguity in the language. According to the theory of "meaning representation from use" proposed by Wittgenstein the various meanings of a word can be represented by other words, and we can assign grades of activation showing the degree of compatibility between labels. A CFS achieves this by using distributions of activations. In a CFS, the activation values agree with the grades of membership and the meaning of a concept is represented by the distribution of the activation values of the other nodes. Because the distribution changes depending on which labels are activated as a result of the conditions the activations show a context-dependent meaning. When more than two labels are activated, a CFS is generated by the overlapping propagations of their activations. In the CFSs, words may have synonymous, antonymous, hypernymous and hyponymous relation to other words. These relations are too complicated to be represented in a hierarchical structure. In this talk, we introduce RBF-like networks to generate CFSs.

Let's think about Java. If we are talking about computers, "java" will be understood as a programming language. If we are looking at a menu at a cafe, it will be understood as a kind of coffee. Its meaning is thus determined by context generated by the presence of related words, such as FORTRAN and C. Experimental results showed that CFSs provide us deferent meaning representations in deferent contexts.

3.1.1 Web community distillation

We applied CFSs to cluster web pages and distill their communities. The applied system processes web pages along with following steps and distills communities of pages. The system is roughly divided into two parts; filtering part (steps 1-6) and classifying part (step 7).

1. Obtain web pages that are similar or linked to a sample web page using Google™.
2. Analyze each HTML file obtained step 1 and generate a word vector.
   - Nouns and adjectives are extracted from the HTML file

- TF-IDF values are calculated and attached to the words

3. Input the word vector into CFSs unit. Propagation of activation occurs from input word vector in the CFSs unit. The meanings of the keywords are represented in other expanded words regarding context.
4. Input the expanded word vector into SVM unit. The SVM unit determines whether the word vector matches to a topic or not, and store the URLs being positive into a database.
5. Repeat steps 2-4 for all HTML files resulted in step 1.
6. Repeat steps 1-5 until there are no new pages found.
7. Classify all web pages in the database and distill communities.

To evaluate this system we selected actual hope pages and simulated community distillation. The results show that conceptual expansion using CFSs has effect to restrain unnecessary words and to emphasize important ones. CFSs also provide us very effective and conceptual measurement performance among text notes.

3.1.2 Image search

We improved Google™ image search capability in two steps as follows.

<Step 1: Relevance feed back>

We developed the system to perform followings relevance feedback circulations. Experimental results show that the relevance feed back improved the image search capability.
- send user's query to Google™
- Google™ sends back retrieved images
- show the images and let the user to select positive examples
- refine query and send it to Google™

<Step 2: Query expansion using CFS>

In the step 1, when popular words are used as a query, such as "cat", relevance feed back did not work well. Popular images are frequently contained in web pages with weak relations with texts in the pages. In this step, CFSs extend query and the results are blended to word vector obtained from analysis of HTML files in the relevance feed back. Experimental results show significantly better results comparing with simple relevance feedback.

Acknowledgements

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

V. Loia, M. Nikravesh and Lotfi A. Zadeh, "Fuzzy Logic an the Internet", to be published in the Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer (August 2003)

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 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.

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

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.

M. Nikravesh, V. Loia,, and B. Azvine, Fuzzy logic and the Internet (FLINT), Internet, World Wide Web, and Search Engines, International Journal of Soft Computing-Special Issue in fuzzy logic and the Internet , 2002

M. Nikravesh, Fuzzy Conceptual-Based Search Engine using Conceptual Semantic Indexing, NAFIPS-FLINT 2002, June 27-29, New Orleans, LA, USA

M. Nikravesh and B. Azvin, Fuzzy Queries, Search, and Decision Support System, International Journal of Soft Computing-Special Issue in fuzzy logic and the Internet , 2002

M. Nikravesh, V. Loia, and B. Azvine, Fuzzy logic and the Internet (FLINT), Internet, World Wide Web, and Search Engines, to be appeared in International Journal of Soft Computing-Special Issue in fuzzy logic and the Internet , 2002

M. Nikravesh, Fuzzy Conceptual-Based Search Engine using Conceptual Semantic Indexing, NAFIPS-FLINT 2002, June 27-29, New Orleans, LA, USA

T. Takagi, et al., Conceptual Fuzzy Sets as a Meaning Representation and their Inductive Construction, International Journal of Intelligent Systems, Vol. 10, 929-945 (1995).

T. Takagi, et al., Multilayered Reasoning by Means of Conceptual Fuzzy Sets, International Journal of Intelligent Systems, Vol. 11, 97-111 (1996).

T. Takagi and M.Tajima, Proposal of a Search Engine based on Conceptual Matching of Text Notes, IEEE International Conference on Fuzzy Systems FUZZ-IEEE'2001, S406- (2001)

T. Takagi, Ket el., Exposure of Illegal Website using Conceptual Fuzzy Sets based Information Filtering System, the North American Fuzzy Information Processing Society - The Special Interest Group on Fuzzy Logic and the Internet NAFIPS-FLINT 2002, 327-332 (2002)

T. Takagi, et al., Conceptual Fuzzy Sets-Based Menu Navigation System for Yahoo!, the North American Fuzzy Information Processing Society - The Special Interest Group on Fuzzy Logic and the Internet NAFIPS-FLINT 2002, 274-279 (2002)

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.

Masoud Nikravesh and Lotfi A Zadeh
Berkeley Initiative in Soft Computing (BISC)
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
CA 94720-1776;
Zadeh@cs.berkeley.edu
Telephone (Zadeh): 510-642-4959; Fax: 510-642-1712
Nikravesh@cs.berkeley.edu
Telephone (Nikravesh): 510-643-4522; Fax: 510-642-5775

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

May 2003
Berkeley, California
USA