

2004

Perception-based fuzzy information retrieval

Hemlata Ganesh Anand Arcot
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Arcot, Hemlata Ganesh Anand, "Perception-based fuzzy information retrieval" (2004). *Master's Theses*. 2596.
DOI: <https://doi.org/10.31979/etd.4fbe-pdwb>
https://scholarworks.sjsu.edu/etd_theses/2596

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

PERCEPTION-BASED FUZZY INFORMATION RETRIEVAL

A Thesis

Presented to

The Faculty of the Department of Computer Science

San Jose State University

In Partial fulfillment

of the Requirements for the Degree

Master of Science

by

Hemlata Ganesh Anand Arcot

May 2004

UMI Number: 1421391

Copyright 2004 by
Arcot, Hemlata Ganesh Anand

All rights reserved.

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 1421391

Copyright 2004 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

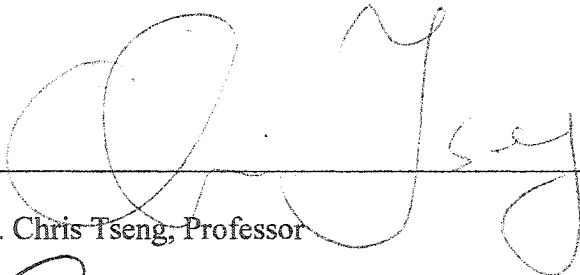
ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© 2004

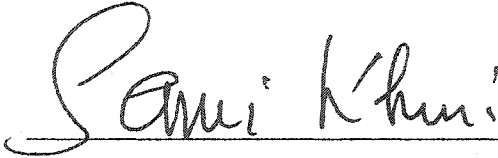
Hemlata Ganesh Anand Arcot

ALL RIGHTS RESERVED

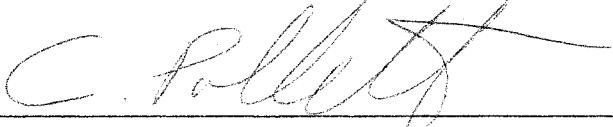
APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE



Dr. Chris Tseng, Professor




Dr. Sami Khuri, Professor



Dr. Chris Pollett, Assistant Professor

APPROVED FOR THE UNIVERSITY



ABSTRACT

PERCEPTION-BASED FUZZY INFORMATION RETRIEVAL

by Hemlata G. Arcot

The World Wide Web today has a wealth of information. However, retrieval of this information using search engines existing today yields imprecise, ineffective, and irrelevant results. In this paper, a novel approach has been proposed to search and rank web pages using fuzzy matching that yields the most relevant results. Fuzzy matching is based on Computational Theory of Perceptions and Fuzzy Logic. The concept is illustrated through the implementation of a “Perception-based Fuzzy Information Retrieval” system that performs “Linguistic-To-Linguistic” and “Linguistic-To-Numeric” searching and ranking. This system has deductive capabilities to conceptually match and rank pages based on predefined linguistic formulations based on human perceptions. In addition to providing a keyword, the user constrains the search by providing a linguistic variable that shall help reduce the number of hits to yield the most relevant results. Illustration of this proposed scheme is provided on the domain of “Food and Health.”

Table of Contents

1	INTRODUCTION	1
1.1	BACKGROUND	1
1.1.1	<i>Search Engines</i>	2
1.1.2	<i>Fuzzy Logic and Computational Theory of Perceptions</i>	5
2	PROBLEM STATEMENT	11
2.1	OBJECTIVE	11
2.2	REQUIREMENTS.....	12
2.2.1	<i>Query Processing</i>	12
2.2.2	<i>Ranking</i>	12
2.2.3	<i>Presentation</i>	13
3	PROBLEM SOLUTION	14
3.1	QUERY PROCESSING	14
3.2	RANKING.....	16
3.2.1	<i>Language Flexibility – Fuzzy Membership Score</i>	17
3.2.2	<i>Relevance – Distance Score</i>	21
3.2.3	<i>Linguistic-To-Linguistic Method - Fuzzy and Distance Score</i>	22
3.2.4	<i>Linguistic-To-Numeric Method - Numeric Score</i>	23
3.3	PRESENTATION.....	25
3.4	FEATURES/FUNCTIONALITY.....	26
3.4.1	<i>User Interface</i>	26
3.5	LIMITATIONS.....	33
4	ARCHITECTURE	34
4.1	DATABASE LAYER	35
4.2	APPLICATION LAYER	35
4.3	PRESENTATION LAYER.....	36
5	DESIGN	37
5.1	MODULE DESCRIPTION	38
5.2	LOGIC FLOW CHART	42
6	EVALUATION OF RESULTS	44
6.1	STANDARD SEARCH VS. FUZZY LINGUISTIC-TO-LINGUISTIC SEARCH.....	44
6.1.1	<i>Local/Domain-Specific Search</i>	45
6.1.2	<i>Meta-Search</i>	47
6.2	STANDARD SEARCH VS. FUZZY LINGUISTIC-TO-NUMERIC SEARCH	49
6.2.1	<i>Local/Domain-Specific Search</i>	50
6.2.2	<i>Meta-Search</i>	52
7	CONCLUSION	55
	REFERENCES.....	56

List of Figures

Figure 1.1 – How Search Engines Work (Sullivan, 2002)	3
Figure 1.2 – Precision and Significance	6
Figure 1.3 – Fuzzy Concept of Tall (Teo, 1995)	7
Figure 1.4 – Crisp Concept of Tall (Teo, 1995)	7
Figure 1.5 – Determining the Degree of Membership in Tall (Teo, 1995)	8
Figure 3.1 – Architectural Diagram of Proposed Solution	15
Figure 3.2 – Flow of Output Presentation Process	25
Figure 3.3 – “Linguistic-To-Linguistic” Search User Interface	26
Figure 3.4 – “Linguistic-To-Linguistic” Search Results	27
Figure 3.5 – “Linguistic-To-Numeric” Search User Interface.....	28
Figure 3.6 – “Linguistic-To-Numeric” Search Results.....	29
Figure 3.7 – “Standard” Search User Interface.....	30
Figure 3.8 – “Standard” Search Results.....	31
Figure 3.9 – “User Preferences”	32
Figure 4.1 – Architecture of Perception-based Fuzzy Meta-Search Engine.....	34
Figure 5.1 – Perception-based Fuzzy Meta-Search Engine Design.....	37
Figure 5.2 – Fuzzy Meta-Search Engine Logic Flowchart	42
Figure 6.1 – Graphical Representation of Standard vs. L2L Search using Local/Domain-Specific Search.....	47
Figure 6.2 – Graphical Representation of the Standard vs. L2L using Meta-Search	49
Figure 6.3 – Graphical Representation of Standard vs. L2N using Local/Domain-Specific Search.....	52
Figure 6.4 – Graphical Representation of Standard vs. L2N using Meta-Search.....	54

List of Tables

Table 3.1	– Sample Entry in Perception Database.....	19
Table 3.2	– Sample Entry in Linguistic-To-Numeric Perception Database	23
Table 6.1	– Standard vs. Fuzzy L2L Search using Local/Domain-Specific Search	46
Table 6.2	– Standard vs. Fuzzy L2L Search using Meta-Search	48
Table 6.3	– Standard vs. Fuzzy L2N Search using Local/Domain-Specific Search.....	51
Table 6.4	– Standard vs. Fuzzy L2N Search using Meta-Search.....	53

1 Introduction

Search engines existing today manage information in a crisp way. The existing search engines try to find an exact match for a query yielding a large set of imprecise and ineffective results. In order to obtain the most relevant results, one has to try to search multiple times by rewording the keywords depending upon the previous results. Moreover, there are only a handful of decent search engines available today that return a good set of URLs (Uniform Resource Locator). The problem is not in the way the search engines return the results but with the highly unorganized and unstructured way of the data on the web. The information available on the World Wide Web (WWW) is highly unstructured; hence, retrieving useful information from the WWW is a challenge faced by most of the search engines existing today. This paper tries to improve the process of information retrieval specifically in the area of relevance by proposing a Perception-based Fuzzy Information Retrieval System.

1.1 Background

This section explains how existing search engines try to tackle the problem of information retrieval and what its drawbacks are. In addition, this section also covers two main concepts of Fuzzy Logic and the Computational Theory of Perceptions since the proposed methodology utilizes these concepts.

1.1.1 Search Engines

There are two basic types of search engines: Crawler-based search engines and human-powered directories (Sullivan, 2002). Crawler-based search engines, such as Google, create their listings automatically. They "crawl" the web, then people search through what they have found. To tackle this unstructured data, the crawler software simply indexes the web pages by keeping a count of each significant word in the web page. Then, every search engine has its own method for ranking the web pages. For instance, in Google, the rank of a web page depends upon the number of web pages that link to it and the ranks of those web pages. In addition, Google also uses a Hypertext-Matching analysis where the entire content of a web page is analyzed to the extent of the font size and location of the words in the web page to ensure that the results returned are most relevant to the user's query. Thus, when one performs a search, the information retrieval occurs based on the indexes built and the ranks assigned to the resultant web pages.

A human-powered directory, such as the Open Directory, depends on humans for its listings. People submit a short description to the directory for their sites, or editors write one (description) for sites they review. A search looks for matches only in the descriptions submitted. There are some search engines such as Yahoo that use both methodologies to optimize the search process.

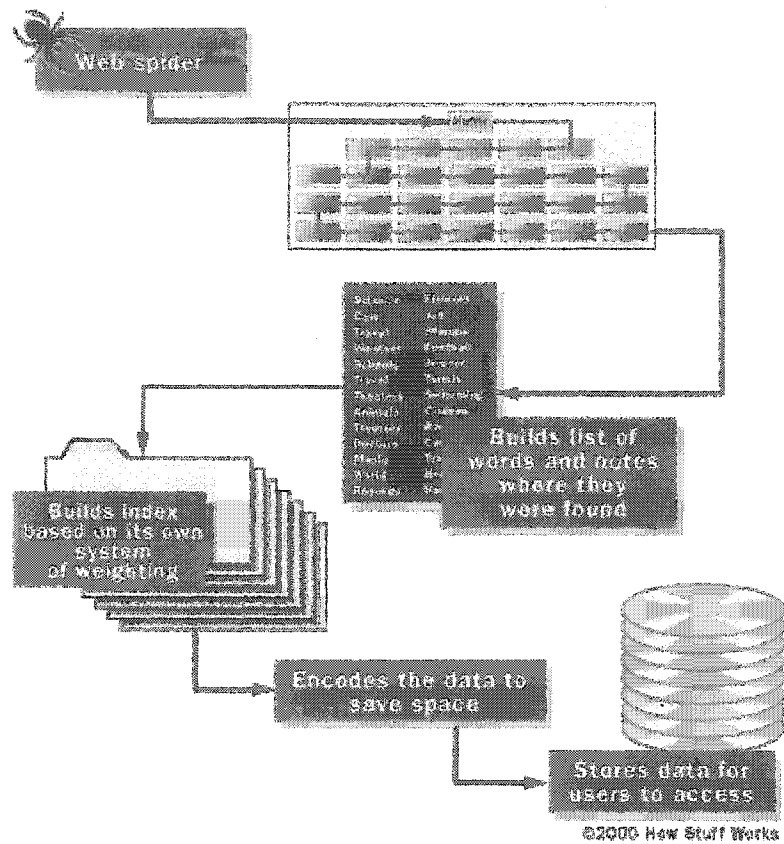


Figure 1.1 – How Search Engines Work (Sullivan, 2002)

Clearly, both techniques have their pros and cons. The information stored by crawler-based search engines is more updated than the information in human-powered directories. However, the crawler software can take days to run depending upon how efficiently it has been programmed. Then, to keep the database up-to-date, the crawler software must be periodically executed. Merely indexing the web pages does not guarantee remarkable results. The major drawback to this scheme is that it lacks intelligence. For instance, the English language is so ambiguous that the same word can be interpreted in a different way and in a different meaning in different contexts. The human-powered directories are good because they have some human intelligence behind them. However, a major

drawback to this scheme is that maintaining such directories require tedious manual work. As the size of the World Wide Web increases, it will become very challenging and difficult for the human-powered directories to keep up with the changes.

Another underlying issue of information retrieval from a user's standpoint is that to get the most relevant results, one may have to query several search engines individually. This can be a frustrating process because every search engine has different usage rules.

However, there is hope for a better future. There is ongoing research in the area of information retrieval from the World Wide Web. There are two specific instances of this work: The Berkeley Initiative in Soft Computing (BISC) for Perception-based Processing and Analysis. The objective of this initiative called FLINT (Fuzzy Logic and the INternet) is to develop an intelligent computer system with deductive capabilities to conceptually match and rank pages based on predefined linguistic formulations and rules defined by experts or based on a set of known homepages (Nikraves, 2002). In effect, the research is trying to apply the Computational Theory of Perceptions to achieve the results. The achievements of this project are unknown at this time.

Another instance of this ongoing research can be found at Stanford University. At Stanford, a group of computer scientists is trying to come up with ways to make web data as machine-readable data instead of human-readable data by creating a TAP (tap.stanford.edu) architecture such that search queries shall produce relevant results (McCool & Guha, 2002). To achieve this goal, they have come up with a concept of semantic-web using RDF (Resource Description Framework). Using RDF, data on a web page is represented as a graph containing resources as nodes and arcs between resources

defining the relationships between those resources. In addition, they are trying to exploit the emerging technology of XML (eXtensible Markup Language) Web-services that use XML syntax and SOAP (Simple Object Access Protocol) protocols to exchange/expose data. While XML defines the syntax of data exchange, the Stanford research group is trying to define guidelines for data *semantics* such that information from different web-services pertaining to a single domain can be integrated to create a knowledge base (McCool & Guha, 2002). This is done by integrating the data-graphs in RDF of each web site into a global graph. In comparison, this thesis has similar goals but a different path. This thesis tries to work with the existing structure of web data. In addition, this thesis is built on the strong foundation of the functioning of a human mind. A human mind is based on perceptions and only perceptions can give a new meaning to the queries to retrieve relevant results. This project uses fuzzy logic as its background technology. While both methods achieve similar goals, both methods have their own shortcomings too. By making human-readable data into machine-readable data, the Stanford's TAP engine loses crucial subjective/perceptive information on the web, which may be important to the users. On the other hand, the proposed methodology in this thesis struggles with semantics of the existing structure of the web data.

1.1.2 Fuzzy Logic and Computational Theory of Perceptions

What is Fuzzy Logic? One thing for sure, Fuzzy logic is not about precision. Henri Matisse (a famous French painter of the 20th century) once said, "Precision is not truth." Dr. Phil (a famous American psychologist) adds, "There is no reality; only perception."

Fuzzy Logic and Computational Theory of Perceptions try to get truth without using precision but by perception.

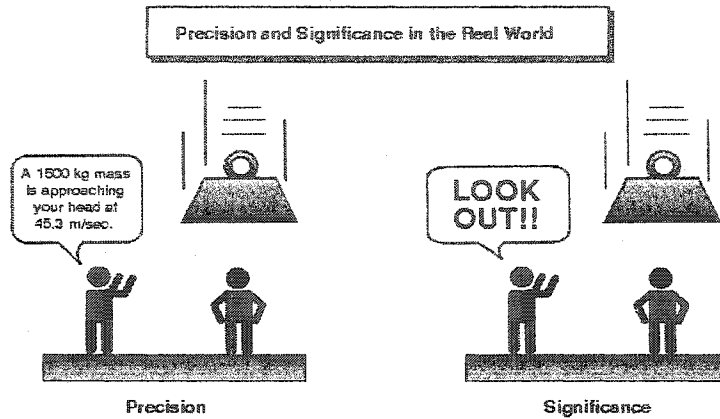


Figure 1.2 – Precision and Significance

Lotfi Zadeh introduced the concept of fuzzy logic in 1965 (Kantrowitz, 1993). After much controversy and skepticism, fuzzy logic found its way into consumer products such as washing machines, camcorders, etc. It is also used in a variety of control applications such as chemical process control and manufacturing. Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth -- truth-values between "completely true" and "completely false". Zadeh of UC/Berkeley introduced this concept in the 1960's as a means to model the uncertainty of natural language. According to Zadeh, "Rather than regarding fuzzy theory as a single theory, we should regard the process of 'fuzzification' as a methodology to generalize ANY specific theory from a crisp (discrete) to a continuous (fuzzy) form" (Kantrowitz, 1993).

In this world, one encounters many things that are imprecise, i.e. they have a certain degree of fuzziness in their description of their nature. An example of a fuzzy statement is *Henry is rather tall but Tom is short*. This statement contains imprecision in

describing height. Then, a question is raised: What height should be considered as tall what height should be considered as short? Fuzzy Logic has an answer to this question of imprecision. The set of heights to be considered tall is not discrete and bounded from those heights considered short (Teo, 1995).

In Figures 1.3 and 1.4, one can see the difference between fuzzy and crisp logic for the human perception of tall.

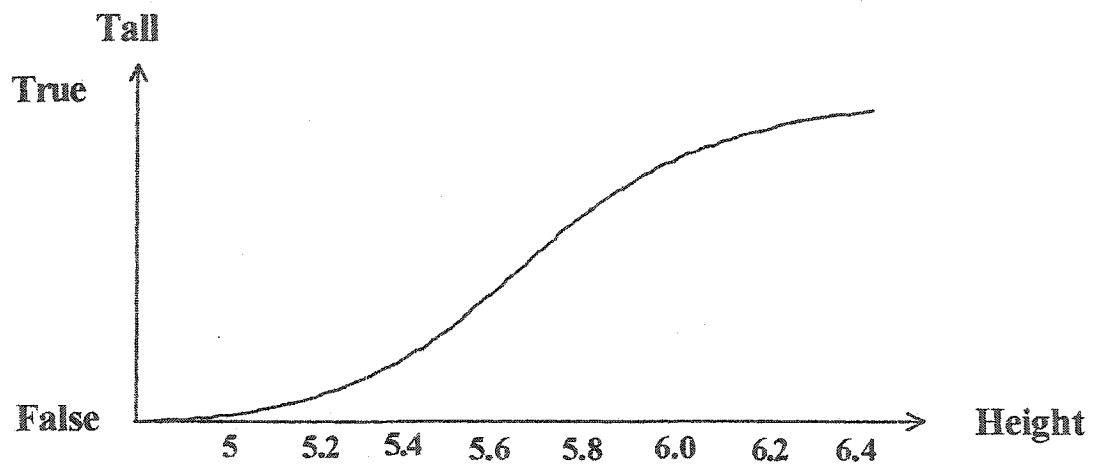


Figure 1.3 – Fuzzy Concept of Tall (Teo, 1995)

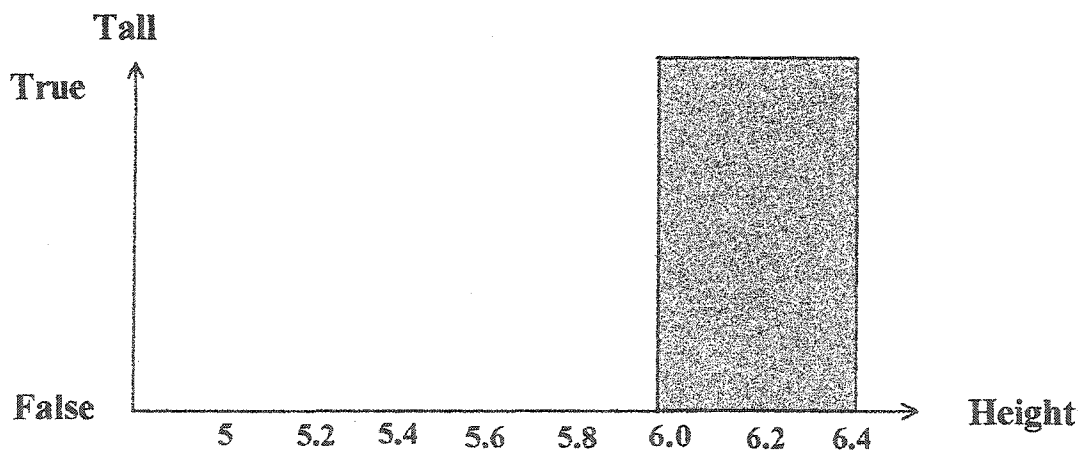


Figure 1.4 – Crisp Concept of Tall (Teo, 1995)

1.1.2.1 Fuzzy set versus crisp sets

As shown in Figure 1.4, the characteristic function for the crisp set reflects its boolean nature. The membership of a height in the set of tall heights remains false (zero) until it reaches exactly six feet when it jumps immediately to true (one). In other words, in crisp sets, an object is either in the set or not in the set.

Fuzzy sets, on the other hand, handle not only true or false values but also regions where it is both true and false. The idea of tall as illustrated in Figure 1.5 is a classical example of a fuzzy set and illustrates the intrinsic properties of fuzzy spaces. The domain of this set, indicated along the horizontal axis, is the range of heights between 5 feet and 6'4". The degree of membership or truth function is indicated on the vertical axis to the far left. In general, the membership goes from zero (no membership) to one (complete membership). The membership function and the domain are connected, in this case, by a simple linear curve; i.e. tallness is directly proportional to height (Teo, 1995). Given a value for height, one can determine its degree of membership in the fuzzy set.

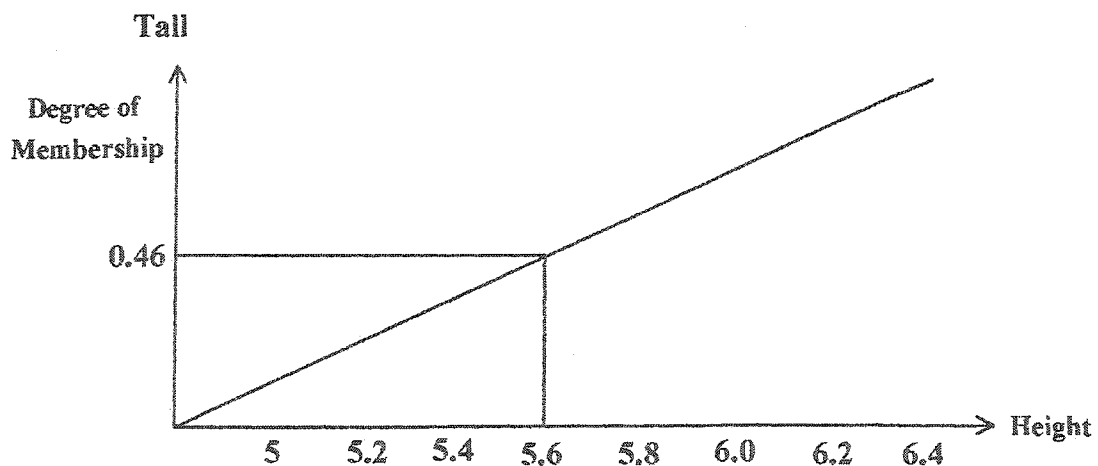


Figure 1.5 – Determining the Degree of Membership in Tall (Teo, 1995)

Thus, a height of 5'6" has 0.46 degree of membership. The interpretation of this value corresponds to the truth of the proposition. If the value of height is less than 5", then its membership is zero. If the height is greater than or equal to 6'4", then its membership value is one. For all values in between 5" and 6'4", the membership value is between 0 and 1 proportionately.

1.1.2.2 What is Computational Theory of Perceptions?

In 1999, Zadeh introduced the Computational Theory of Perceptions. In Zadeh's words, "Humans have a remarkable ability to perform a wide variety of physical and mental tasks without any measurements and any computations. Familiar examples of such tasks are parking a car, driving in heavy traffic, playing golf, riding a bicycle, understanding speeches and summarizing a story. Underlying this remarkable ability is the brain's crucial ability to manipulate **perceptions** - **perceptions** of distance, size, weight, color, speed, direction, force, numbers, truth, likelihood and other characteristics of physical and mental objects." Thus, according to Zadeh, manipulation of **perceptions** plays a key role in human recognition, decision, and execution processes. The methodology of Computing with Words (CW) provides a foundation for **Computational Theory of Perceptions** - a theory that may have an important bearing on how humans make - and machines might make - rational decisions in an environment of imprecision, uncertainty, and partial truth (Zadeh, 1999).

In Computational Theory of Perceptions, semantic representation is based on what is referred to as Constraint-centered Semantics of Natural Languages (CSNL). The theory

assumes that perceptions are described by propositions (declarative phrases) drawn from a natural language. A proposition, p , is viewed as an answer to a question and the meaning of p is represented as a generalized constraint. For example, the propositional phrase *bright flowers* answers the question of what kind of flowers, and the word *bright* is the generalized constraint that represents the perception of color. To compute with perceptions, their descriptors are translated into what is called the Generalized Constraint Language (GCL) (Zadeh, 1999).

The proposed thesis strives to apply Computational Theory of Perceptions to the problem of search and retrieval of web data. To create an intelligent retrieval system, a Perception-based Fuzzy Meta-Search Engine has been proposed. The system is capable of querying multiple search engines for results and it incorporates a perception index to add deductive capabilities. Querying multiple search engines takes best advantage of both kinds of search engines, human-powered directories and crawler-based engines. In addition, this engine uses a multitude of fuzzy terms to match up with the keywords being searched. This adds the human perception element to obtain a relevant set of results. The result set can be reused or integrated very easily as it is generated using standard XML methodologies.

2 Problem Statement

This chapter defines the objective and states the requirements for a Perception-Based Fuzzy Meta-Search Engine. The requirements are presented in three major areas of query processing, ranking, and presentation.

2.1 Objective

The goal of this project is to improve the information retrieval process by creating a perception-based information processing and retrieval system. The objective is to develop a **“Perception-based Fuzzy Meta-Search Engine”** with deductive capabilities to conceptually match and rank pages based on linguistic formulations. In other words, users searching the web can narrow thousands of hits to the few that they really want by using fuzzy terms in their queries. The fuzzy terms shall represent human perceptions of time, direction, speed, shape, and many other attributes of physical and mental objects; hence, the name, perception-based information retrieval. The human-like query system will present the search results based on how close in meaning they are, to the provided keywords. In addition, the system shall query search engines like Google, Yahoo, AllTheWeb, MSN, and AltaVista to present a consolidated set of results in a convenient summarized format using standard XML formats.

2.2 Requirements

2.2.1 Query Processing

The Perception-based Fuzzy Meta-Search Engine shall improve the search process by querying multiple search engines at once. Hence, the system shall allow users to select search engines, which they wish to be queried, from a pre-defined set of search engines.

1. The system shall format the input search keywords to each search engine's requirements.
2. The system shall allow search for "all words" or "any of the words" or "exact phrase" in the keywords provided.
3. It shall also extract and present only relevant and distinct search result links from each search engine. Thus, the system should be capable of identifying and eliminating sponsored results and advertisements returned by each search engine.

2.2.2 Ranking

The Perception-based Fuzzy Meta-Search Engine shall improve the search process by ranking the results obtained from multiple search engines.

1. Ranking shall incorporate some deductive capabilities such as the most relevant results receive a higher rank. In order to have deductive capabilities in the Perception-based Fuzzy Meta-Search Engine, it should possess the behavior of a human mind to some extent. When users view the results of a search, they have to filter the results further to pick the URLs that seem more relevant and appealing.

Implementation of this step in the Perception-based Fuzzy Meta-Search Engine will vastly improve the search process by making it more efficient and satisfying.

2. The Perception-based Fuzzy Meta-Search Engine should overcome the limitations of imprecision and vagueness since the web has a wealth of information in a highly unstructured form. Overcoming these limitations shall help control the size of the results, express soft retrieval conditions, and produce a discriminated (precise) set of results (Choi, 2003).
3. Existing search engines do not do a good job of modeling human perceptions. Modeling human perceptions are key to an intelligent Perception-based Fuzzy Meta-Search Engine. Human perceptions are fuzzy in nature because of the way humans interpret a language and understand the ambiguity within a language.

2.2.3 Presentation

The Perception-based Fuzzy Meta-Search Engine shall present consolidated search results in a standard summarized format.

1. It shall keep the content separate from the presentation so that any changes to presentation shall not require modifications to the Perception-based Fuzzy Meta-Search Engine.
2. It shall present descriptive information of each resultant URL for the title, description, keywords, etc.
3. The resultant URL description shall contain the fuzzy terms (if any) and keywords highlighted so that user can quickly identify the regions of interest.

3 Problem Solution

The following sections discuss how each requirement is addressed and implemented. In addition, the chapter covers the methodologies used to solve each problem and the rationale for selecting those methodologies. Figure 3.1 below visualizes the steps followed to accomplish the results. Later, in Section 3.4, it also includes some screenshots of the user interfaces of the Perception-based Fuzzy Meta-Search Engine. The chapter concludes with a list of limitations of the Perception-based Fuzzy Meta-Search Engine.

3.1 Query Processing

The Perception-based Fuzzy Meta-Search Engine shall process the query in the following ways:

1. **Input Formatting:** The Perception-based Fuzzy Meta-Search Engine translates query keywords into the necessary format as expected by each search engine. It submits the translated query to multiple search engines.
2. **Information Retrieval:** The Perception-based Fuzzy Meta-Search Engine opens an HTTP (Hyper Text Transfer Protocol) connection to each search engine web page and appends the keyword to the URL as a query string. Most Search engines use the HTTP “get” method to retrieve the input keyword. For instance, a search to Yahoo can be processed by providing the keyword in the URL as <http://search.yahoo.com/bin/search?p=food> where “food” is the input keyword to be searched. As of today, the query is submitted to Google, Yahoo, AllTheWeb,

AltaVista, and MSN Search. The engine also allows the user to choose whether to search for “all the words” or “any of the words” or “exact phrase” among the keywords he/she provides. This option is directly passed to each search engine in the URL such that the request is fulfilled accurately.

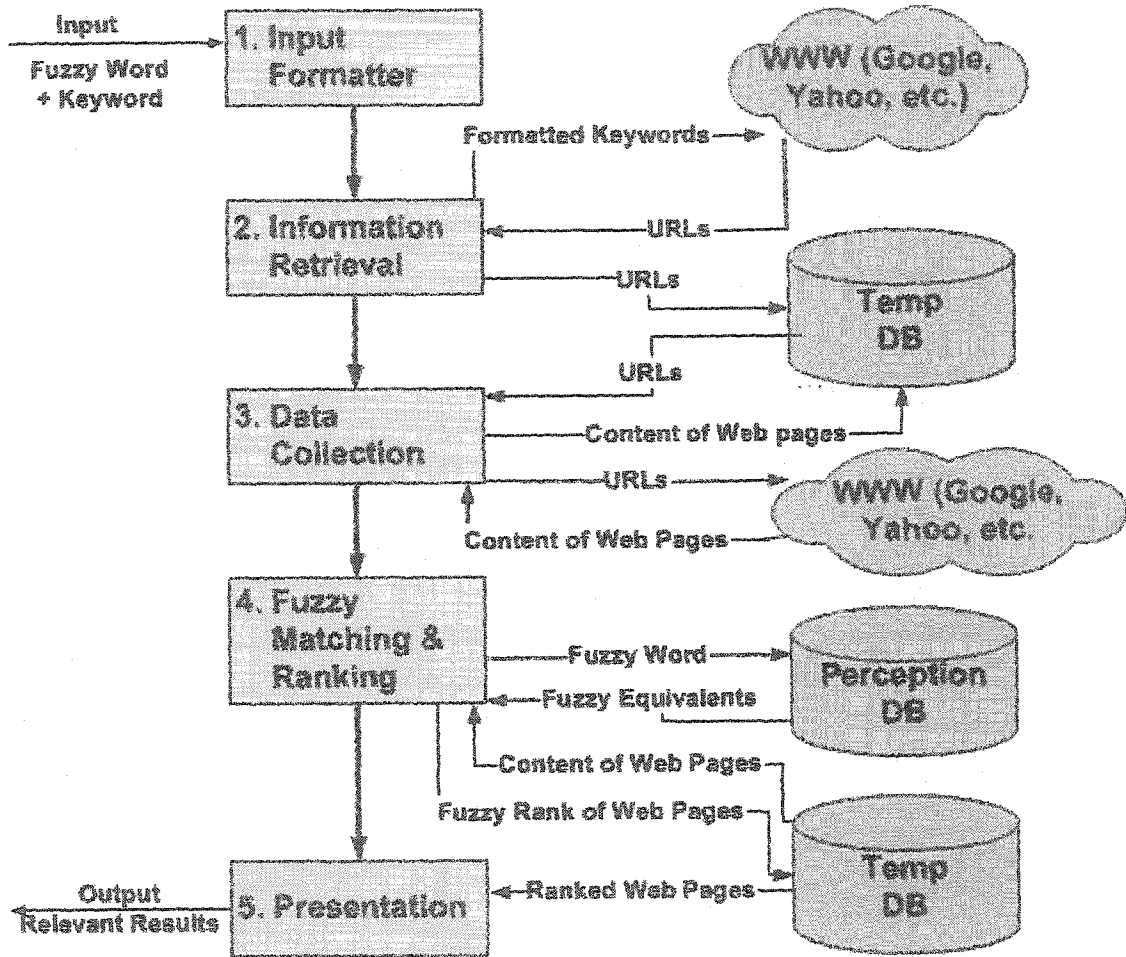


Figure 3.1 – Architectural Diagram of Proposed Solution

3. **Information Processing:** The results from each search engine are then parsed to extract the links using a third party API (Application Programming Interface) called “HTMLParser.” HTMLParser is a Java API that takes a URL, opens a connection to

the URL, and loads the entire web page into a collection of HTML (Hyper Text Markup Language) nodes. The nodes can be parsed to extract any kind of content information such as links or images or text, etc. Irrelevant links such as advertisements and directory matches are eliminated through some customization.

4. **Duplicate Elimination:** After retrieving all resultant URLs from each search engine, the Perception-based Fuzzy Meta-Search Engine eliminates the duplicate URLs returned by different search engines. The URLs are stored in a temporary table. Only distinct URLs are picked for further processing to eliminate duplicates. However, the engine does keep track of the source search engine of each resultant URL.
5. **Data Collection:** The Perception-based Fuzzy Meta-Search Engine then visits each resultant link to collect information such as Title, Description, Keywords, and Body of that web page. This information shall be later used for fuzzy matching, to further eliminate irrelevant results, and to present the results.

3.2 Ranking

The goal of this project is to retrieve the most relevant results. To facilitate this objective, an innovative ranking algorithm has been proposed and implemented to rank the web pages in an intelligent manner. This algorithm uses three major criteria to deduce and rank pages. The criteria used are as follows: (a) Language flexibility (Fuzzy membership score), (b) Relevance (Distance Score), and (c) Numeric relevance (Numeric

Score). Using these criteria, two major search and ranking methods are proposed: (a) Linguistic-To-Linguistic Method, and (b) Linguistic-To-Numeric Method.

3.2.1 Language Flexibility – Fuzzy Membership Score

To facilitate the goal of retrieving the most relevant results, it is proposed that the user provide a **fuzzy term** in addition the **keyword** being searched. For example, the search input “high cholesterol” contains the fuzzy term “high” and keyword “cholesterol.” For clarification purposes, the **keyword** can be any crisp term that bears no ambiguity and represents crisp information. The **fuzzy term** provided by the user should depict a certain human perception. A fuzzy term is imprecise, ambiguous, and is highly subjective in nature. For instance, the term “small” may have different meaning based on the context. It may also have different meanings to different people. The word “small” may refer to a perception of “size” or “amount.” It will be used to filter and rank the results. This idea of introducing a fuzzy term along with the keyword has been proposed by Choi (2003) in a research article, “Enhancing the power of web search engines by means of fuzzy query,” published in “Decision Support Systems.”

To illustrate how a fuzzy term helps, it is important to understand what the data looks like. The data is unstructured web data in natural language. If the ranking engine has to give the most relevant results, it should possess capabilities to handle natural language data in a specific way to perform computations.

According to the Computational Theory of Perceptions, natural languages may be viewed as systems for describing perceptions. Since perceptions are nothing but words,

in effect, computing and reasoning with perceptions is reduced to computing and reasoning with words (Zadeh, 1999). This methodology is referred to as Computing with Words (CW) (Nikravesh, 2002). CW is the backbone of the Computational Theory of Perceptions. To be able to compute with perceptions, it is necessary to have a means of representing their meaning such that it is possible to perform computations on them. The aim is to develop an automated capability to reason with perception-based information. As discussed in Section 1.1.2, Zadeh (1999) proposes the Generalized Constraint Language describing perceptions to give meaning to the data.

So, in this project, this concept of perception-based information processing from the Computational Theory of Perceptions is borrowed to process data in web pages. Using this concept, similar words that depict a perception are grouped and many such groups are used to create a **perception database**. The words are ordered within a group such that words with similar meaning are placed closer together and assigned a number called the score (or value). The score is just a sequential number from 1 to n such that two words that are closer in meaning are closer together in the sequence. For instance, the words “low” and “less” are closer to each other in meaning compared to “tiny” as shown below in Table 3.2. Hence, they have a number of 1 and 2 respectively compared to 4 for “tiny.” This score will be used to rank the web pages based on availability of a word in the content of the web page and the word’s vicinity (in meaning) to the actual fuzzy word provided by the user. The equation 3.1 as shown below has been chosen to apply fuzzy logic for fuzzification of scores/values assigned to each fuzzy term. In fuzzy logic, one can use a set of if-then-else rules or a fuzzy graph (rather its formula) to represent the

rules. Here, a fuzzy graph has been chosen that calculates the proximity between the fuzzy terms in a perception group. In other words, this formula determines whether or not “tiny” is closer to “less” than “low”. So, the fuzzy membership is calculated based on the formula:

$$\text{Membership} = e^{-\frac{(x - \bar{x})^2}{\sigma}} \text{-----(3.1)}$$

where x = Score/value of one of the fuzzy terms in the perception group,

\bar{x} = Score/value of the fuzzy term being queried, and

σ = Sigma (width factor of the fuzzy membership function)

For instance, the perception database may contain the following sample data:

Abstract Perceptions	Perceptions	Fuzzy Terms	Score/Value
Size/Amount	Small	Low	1
		Less	2
		Small	3
		Tiny	4
		Trace	5
Size/Amount	large	High	1
		More	2
		Large	3
		Substantial	4
		Significant	5
		Healthy	6
Size/Amount	medium	Some	1
		Good	2
		Medium	3
		Moderate	4
		Fair	5
		Decent	6

Table 3.1 – Sample Entry in Perception Database

In this example, if the user provides a fuzzy term “less” in the query, and if the current web page being ranked contains the word “low” near the keyword, then the fuzzy membership of the web page is calculated as per the above formula as follows:

$$\text{Membership} = e^{-\frac{(1-2)^2}{5}} = 0.8187 \quad (\text{Given } \sigma = 5)$$

By providing a fuzzy term that constrains a keyword being searched, in effect, the user is using the Generalized Constraint language. GCL provides language flexibility. Hence, the fuzzy meta-search engine requires the user to provide a fuzzy term that constrains the keyword being searched.

The proposed methodology works only on a specific domain because the perceptions and their weights change from one domain to another. For instance, the perception of size can change from a domain of airplanes to dress sizes. So, to demonstrate the proposed methodology, a domain of food and health has been used. The reason for choosing food and health related domain is because it is a general subject and it would be easy to demonstrate the system’s capability using a general topic, and the perception database shall have fuzzy values most appropriate for food and health related topics.

Now, the ranking engine uses this “food and health” perception database as one of the steps to rank web pages. So, when the user provides a fuzzy term, the ranking engine parses each web page for the fuzzy term provided by the user and its equivalents as found in the perception group. The terms are pre-ordered by their meanings and their score/value gives a measure of their closeness in meaning. For example, the term “small” is more closer to “Less” and “Tiny” and less closer to “Low” and “Trace.” Here, fuzzy

logic is chosen instead of boolean logic to allow all of the terms in the group to have some significance such that web pages containing any of these terms will be considered good candidates for relevance. Thus, the engine applies fuzzy rules in the form of a membership function to determine the membership of each fuzzy term in the group when compared to the fuzzy term being queried by the user. The notion of relevance is directly proportional to the degree of membership. The higher the membership of a term in a group, the more relevant it is in terms of ranking. Thus, each term in the group has a fuzzy membership score depicting its relevance.

3.2.2 Relevance – Distance Score

The ranking of a web page is determined by the existence of a fuzzy variable (fuzzy term and its equivalents in the perception group) in its content and the proximity of the fuzzy variable to the keyword. In addition to allowing for language flexibility, it is also desired that the fuzzy term constraining the keyword should really do so in the content of the result to be considered relevant. Mere presence of the fuzzy term and the keyword in the same web page does not guarantee a relevant result. Here, a relevant result is guaranteed only when the fuzzy term is directly next to or in close vicinity of the keyword. A value for distance is computed between the fuzzy term and the keyword in a systematic way. The textual web-data is broken down into tokens where each token corresponds to a sentence. A sentence for computational definition is a sequence of words that end with a period (.) or a semi-colon (;) or a question mark (?) or an exclamation mark (!). Using this rule, the web data is broken down into sentences and

then the distance is computed between the fuzzy term and the keyword. The distance score is arrived at based on the following criteria:

1. When the fuzzy term is adjacent to the keyword, it is considered as an ideal match and that web page gets the highest possible distance score.
2. When the fuzzy term and keyword are in the same sentence, it is given a good but slightly lower distance score because of the nature of the language. This is based on the assumption that a fuzzy term and the keyword in the same sentence also have a certain degree of relevance.
3. When the fuzzy term and the keyword are on the same web page, a small distance score is allocated.

3.2.3 Linguistic-To-Linguistic Method - Fuzzy and Distance Score

The fuzzy score and the distance score are combined (multiplied) to give an overall relevance score to each web page. This is the “defuzzification” step in the application of fuzzy logic where multiple outputs are then combined to give one single output. Here, the method of algebraic product of memberships is chosen to get the final membership (or rank) value. This search and ranking method is called “Linguistic-To-Linguistic” method because a linguistic fuzzy term is mapped to several other linguistic terms depicting the same perception. The higher the relevance score, the more relevant the given web page to the user’s query.

3.2.4 Linguistic-To-Numeric Method - Numeric Score

In addition, the Perception-based Fuzzy Meta-Search Engine has a framework to create and implement fuzzy membership rules and intelligent correlations between perceptions related to the food and health domain and their measurable values such that a fuzzy linguistic variable is mapped to a numeric value. For instance, the perception of “resourcefulness” of a particular nutrient, say “Vitamin A,” in a food has 3 distinct fuzzy linguistic variables such as “low”, “medium”, and “high.” Vitamin A is measured in International units (IU). An adult needs about 4000 to 5000 IU of Vitamin A every day. No single serving of any food has the entire daily value of Vitamin A. So, a food containing at least 25% of the daily value for Vitamin A is considered as “high” in Vitamin A. Similarly, those foods with less Vitamin A are considered as moderate or poor sources according to pre-defined numeric percentages. Using this method, a perception database is created that contains daily value for each ingredient and percentages depicting the resourcefulness of each ingredient. Some sample entries in the database are:

Ingredient	Daily Value	Perception of resourcefulness	Percentage of Daily Value
Vitamin A	5000	High	25%
		Medium	15%
		Low	8%

Table 3.2 – Sample Entry in Linguistic-To-Numeric Perception Database

Given this information in a database, when the fuzzy linguistic variable such as “high” is used in a query containing a keyword such as “Vitamin A,” then the meta-search engine parses each web page for a numeric value of that perception (such as Vitamin A)

and calculates a fuzzy membership value for the actual value in the web page according to its position in the pre-defined fuzzy definitions (% of daily value). It ranks the web pages according to the fuzzy membership of actual value in the pre-defined fuzzy set. The equation 3.2 as shown below has been chosen to apply fuzzy logic for fuzzification of expected and actual values. In fuzzy logic, one can use a set of if-then-else rules or a fuzzy graph (rather it's formula) to represent the rules. Here, a fuzzy graph that calculates the proximity between the expected and actual values has been used. In other words, this formula determines whether or not 1300 IU of Vitamin A can be classified as "high" amount of Vitamin A.

The fuzzy membership is calculated based on the formula:

$$\text{Membership} = e^{-\frac{(x - \bar{x})^2}{\sigma}} \text{-----(3.2)}$$

where x = Actual value of the perception as found in the web page,

\bar{x} = ideal value of the perception as set in the percentage of daily value of the vitamin, and

σ = Sigma (width factor of the fuzzy membership function)

So, here for instance, if the user queries for "high vitamin a," then the fuzzy membership for the web page containing a value of 1200 IU for the vitamin a will be as follows:

$$\text{Membership} = e^{-\frac{(1200 - (5000 * 25\%))^2}{250000}} = 0.99 \quad (\text{Given that } \sigma = 250000)$$

3.3 Presentation

The results of the search and ranking process are stored in a temporary database that exists only as long as the page is loading. The Presentation module converts these results into XML format. XML methodology has been used to present the resulting set of URLs in an agreed upon standard format. XML coupled with XSL (XML Style sheets) makes it possible to keep the content separate from the presentation. Therefore, it is easy to maintain the software. All presentation related changes are done to an XSL document that is interpreted by Internet Explorer to display the results in HTML.

1. The Presentation module consolidates the descriptive information of each resultant URL for the title, description, keywords, etc.
2. The resultant URL description contains the fuzzy terms (if any) highlighted in pink and keywords highlighted in yellow so that user can quickly identify the regions of interest.
3. Fuzzy scores, distance scores, and relevance scores can be displayed by choosing the selections in the preferences section of the search engine.

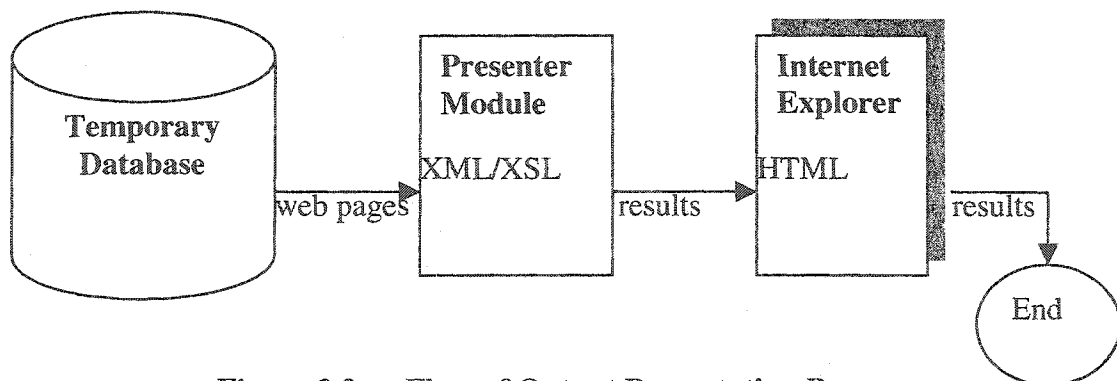


Figure 3.2 – Flow of Output Presentation Process

3.4 Features/Functionality

3.4.1 User Interface

3.4.1.1 Linguistic-To-Linguistic Search

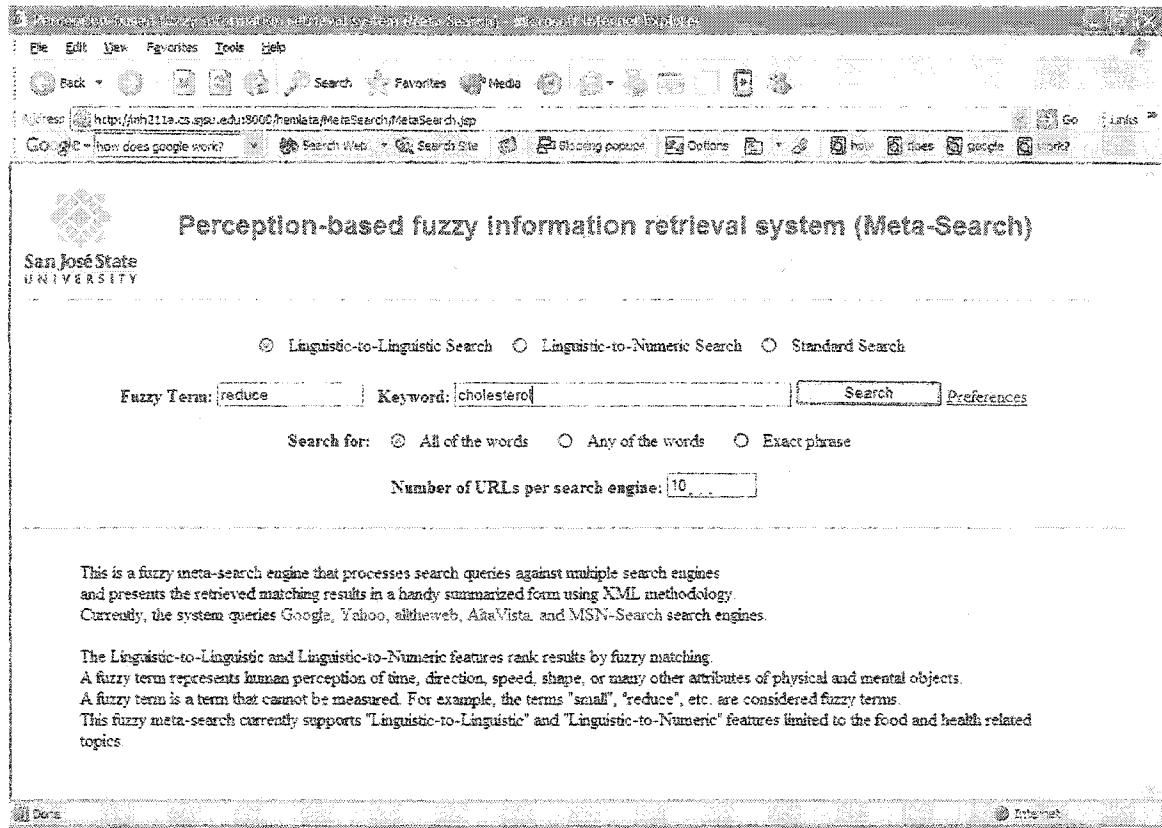


Figure 3.3 – “Linguistic-To-Linguistic” Search User Interface

3.4.1.2 Linguistic-To-Numeric Search

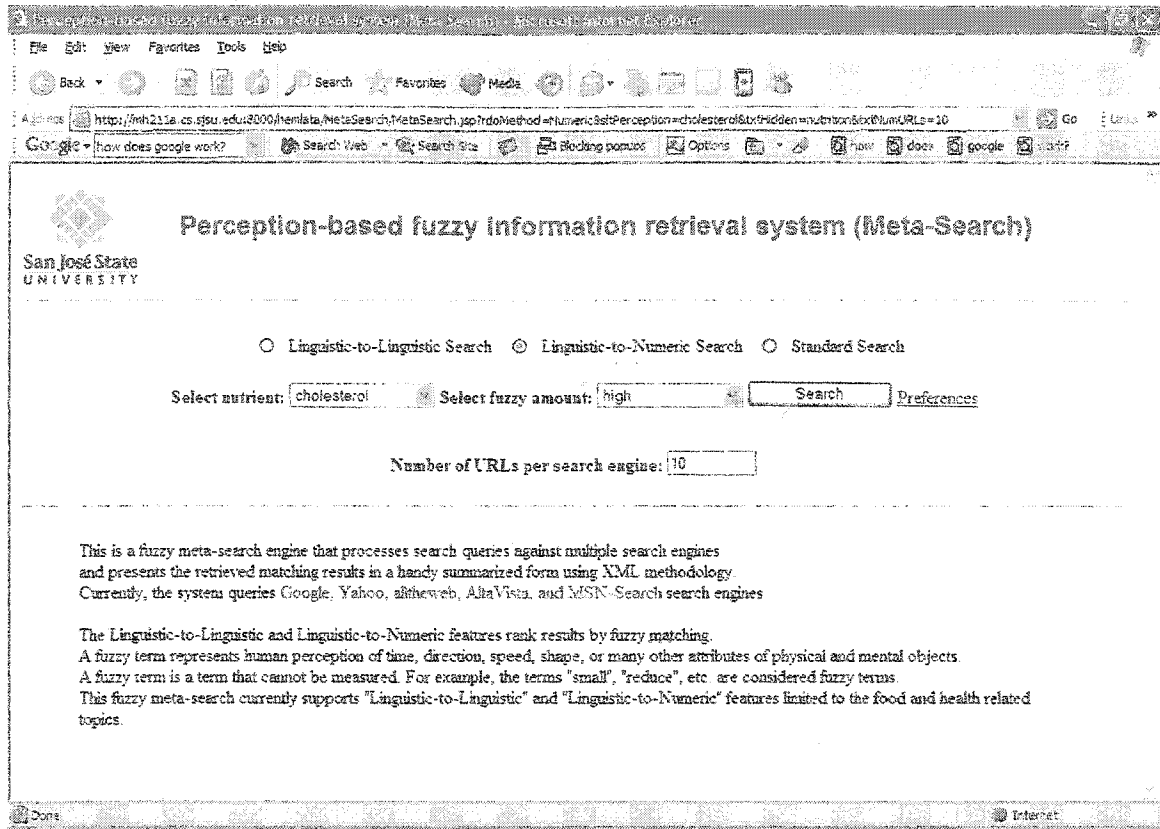


Figure 3.5 – “Linguistic-To-Numeric” Search User Interface

Perception-based fuzzy information retrieval system (Meta-Search)

San José State UNIVERSITY

Search Method: Linguistic-To-Numeric Search Terms: cholesterol
 Search Fuzzy Terms: high Fuzzy Group: high (75); medium (45); low (24)
 Number of Results: 22

Note: The fuzzy scores are on a scale of 0 to 1. If you had provided a fuzzy term, then the "Fuzzy Terms" section shall display all terms for which fuzzy matching was done.

Fats amp; Cholesterol: Nutrition Source, Harvard School of Public Health
 Sources: Yahoo! AltaVista MSN Search Google
 Numeric Value: Score: 1
 Description: harvard school of public health and nutrition research covers latest information on fiber fats calcium carbohydrates eggs nutritional pyramids and fruits and vegetables
 Keywords: harvard school of public health nutrition nutrition research nutrition information nutrition myths fiber colon cancer fats saturated fat unsaturated fat trans fat calcium carbohydrates complex carbohydrates eggs nutritional pyramid fruit
 Body: cholesterol less than milligrams per deciliter (mg/dl) hdl cholesterol levels greater than 40 mg/dl ldl cholesterol levels less than 100 mg/dl dietary fat dietary cholesterol and blood cholesterol levels one

Fats amp; Cholesterol: Nutrition Source, Harvard School of Public Health
 Sources: Yahoo!
 Numeric Value: Score: 1
 Description: harvard school of public health and nutrition research covers latest information on fiber fats calcium carbohydrates eggs nutritional pyramids and fruits and vegetables
 Keywords: harvard school of public health nutrition nutrition research nutrition information nutrition myths fiber colon cancer fats saturated fat unsaturated fat trans fat calcium carbohydrates complex carbohydrates eggs nutritional pyramid fruit
 Body: cholesterol less than milligrams per deciliter (mg/dl) hdl cholesterol levels greater than 40 mg/dl ldl cholesterol levels less than 100 mg/dl dietary fat dietary cholesterol and blood cholesterol levels one

Nutrition Fact Sheet: Dietary Cholesterol, Nutrition, Feinberg School of Medicine

Figure 3.6 – “Linguistic-To-Numeric” Search Results

3.4.1.3 Standard Search

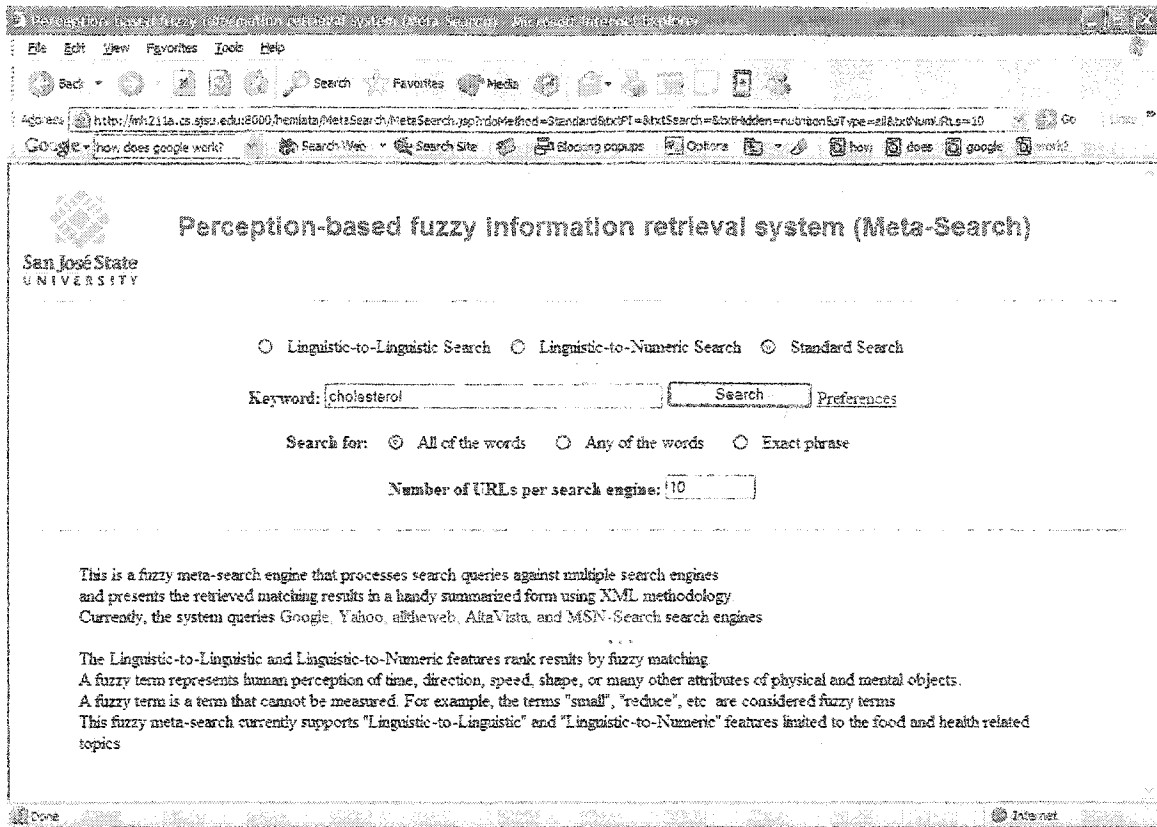



Figure 3.7 – “Standard” Search User Interface

Perception-based fuzzy information retrieval system (Meta-Search) Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://web211a.cs.sjsu.edu/S000/kenata/PerceptionBasedFuzzyInformationRetrievalSystem/MetaSearch/MetaSearchResults.jsp?doMethod=Standard&doSearch=cholesterol&doHidden=nutrition&doType=all&doNumURL

Google how does google work? Search Web Search Site Blocking popups Options how does google link?

 **Perception-based fuzzy information retrieval system (Meta-Search)**

San José State UNIVERSITY [New Search](#)

Search Method: Standard Search Terms: cholesterol

Number of Results: 44

Note: The fuzzy scores are on a scale of 0 to 1. If you had provided a fuzzy term, then the "Fuzzy Terms" section shall display all terms for which fuzzy matching was done.

The Cholesterol Myths

Sources: Yahoo, AlltheWeb, AltaVista, MSN Search, Google

Description:

Keywords: cholesterol, kolesterol, saturated fatty acids, fatty acids, polyunsaturated fatty acids, animal fat, vegetable oils, coronary heart disease, atherosclerosis, ariat, infection, heart, heart attack, heart disease, ldl, hdl, low chole

Body: cholesterol myths the cholesterol myths

Elevated Cholesterol: Doctor's Guide to the Internet

Sources: Yahoo, AlltheWeb, MSN Search, Google

Description: the latest medical news and information for patients or friends/parents of patients diagnosed with elevated cholesterol and elevated cholesterol-related disorder

Keywords: cholesterol, lipitor, cholestagen, lovastatin, hdl, zocor, pravastatin, lescol, hormone replacement therapy, bad cholesterol, good cholesterol, pravachol, soy, t, mevacor, fluvastatin, estrogen, women's health, medical news, health n

Body: cholesterol doctor's guide to the internet

NHLBI NCEP Cholesterol Counts for Everyone Page

Sources: Yahoo, AlltheWeb, AltaVista, Google

Description: this web site contains professional and general health related information on cholesterol lowering. the national heart, lung, and blood institute is a part of the u

Keywords: cholesterol, heart disease, hdl, ldl, coronary heart disease, chd, medicines, drugs, diet, prevention, overweight, physical activity, fat, saturated fat, triglycerides, lung, and blood institute

Body: cholesterol counts for everyone page please provide us your feedback! | updated atp iii guide

Done Internet

Figure 3.8 – “Standard” Search Results

3.4.1.4 User Preferences

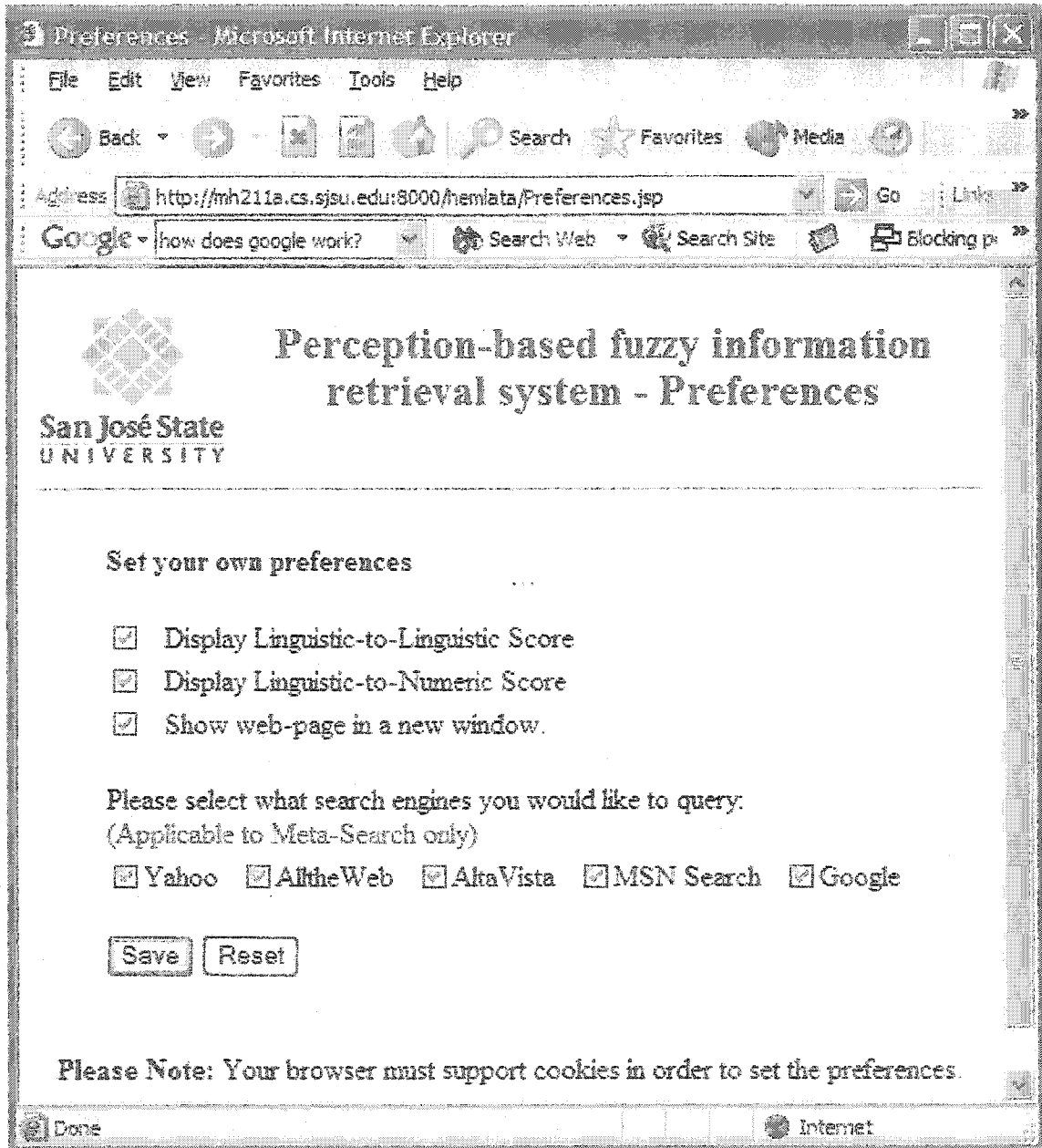


Figure 3.9 – “User Preferences”

3.5 Limitations

Following are the limitations of the Perception-based Fuzzy Meta-Search Engine:

1. Each search engine may actually return maximum of 50-100 URLs per page. So, the Perception-based Fuzzy Meta-Search Engine does not parse beyond the first result page.
2. Standard search results (non-fuzzy) are not ranked in any way.
3. The Perception-based Fuzzy Meta-Search Engine cannot parse all URLs. Some web-servers have high security such that URLs cannot be accessed in any other way other than a browser using HTTP protocol / HTTPS protocol.
4. Non-English language web pages may not get parsed/presented correctly.
5. Web documents in formats other than HTML may not be parsed. For instance, the implemented Meta-Search engine cannot parse PDF documents.

4 Architecture

A 3-tier architecture was used to address the functionality of the Perception-based Fuzzy Meta-Search Engine.

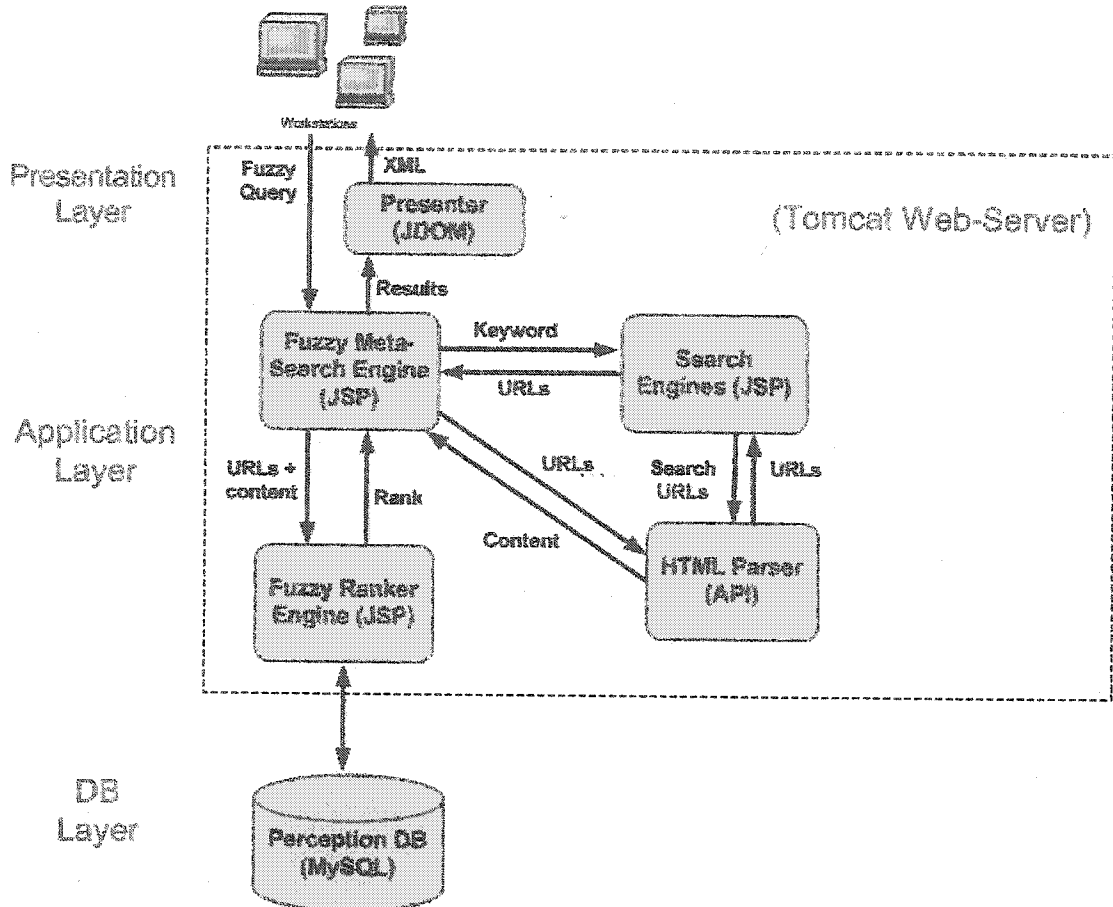


Figure 4.1 – Architecture of Perception-based Fuzzy Meta-Search Engine

4.1 Database Layer

The database layer comprises of the following:

1. The back-end database is MySQL v 3.23.55.
2. The MySQL stores the perception database that contains groups of fuzzy terms pertaining to perceptions of food and health domain. The fuzzy terms are scored based on their closeness to other fuzzy terms in a group.
3. The perception database design is shown in the design section of this document.

4.2 Application Layer

The application layer has been developed using JSP (JavaServer Pages) v2.0 hosted on Apache Tomcat 4.1.18 web-server. The application layer consists of a Perception-based Fuzzy Meta-Search Engine that interacts with different modules to meet the requirements. The Perception-based Fuzzy Meta-Search Engine interacts with the following modules:

1. **Search Engine:** The search engine takes the input keywords and returns a list of URLs and their corresponding details for that keyword. It queries other search engines such as Google, Yahoo, AllTheWeb, MSN, and Altavista to retrieve search results. The search engine uses a JAVA package called HTMLParser v 03302003 to parse other search engine results and extract the web page content.
2. **Page Ranker:** The Page ranker takes the URLs and their content details and ranks the URL based on combination of fuzzy score and distance score or linguistic-to-numeric

membership score. The page ranker utilizes the JFLEX lexer/scanner to break the web page content into sentences and scan it for fuzzy information.

3. **Presenter:** The presenter module collects the ranked search results and outputs the results using XML and XSL. It uses JDOM version beta 8 to create an XML DOM (Document Object Model) document containing search results to be displayed.

4.3 Presentation Layer

The user interface to the Perception-based Fuzzy Meta-Search Engine is an Internet Explorer browser window of version 5.5 and above. The query interface and the results interface are like that of any search engine. The query interface allows user to provide any fuzzy term and a keyword. The fuzzy term shall restrict the keyword. The results contain a list of URLs and their descriptions. The results also contain what other fuzzy terms were queried for the fuzzy term provided by the user.

5 Design

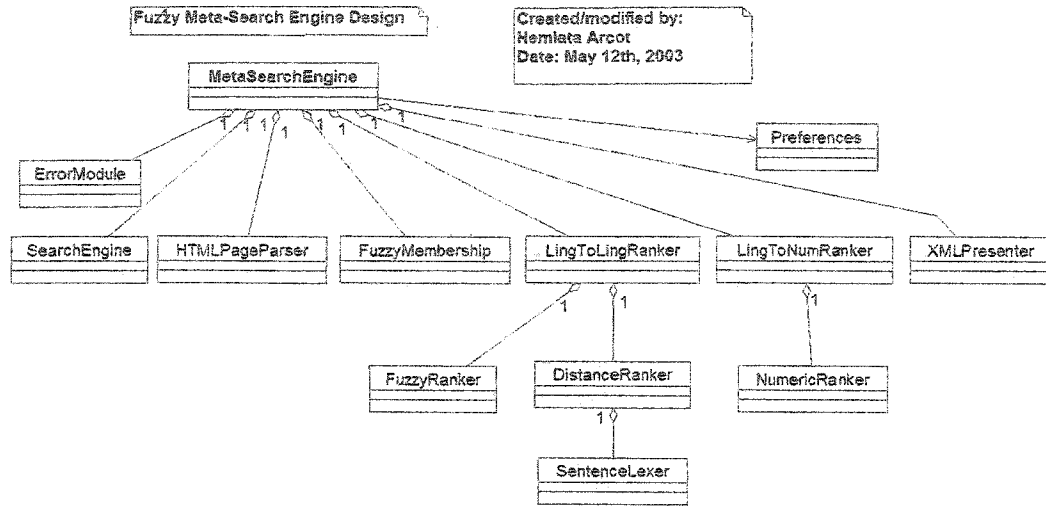


Figure 5.1 – Perception-based Fuzzy Meta-Search Engine Design

5.1 Module Description

At the heart of the Perception-based Fuzzy Meta-Search Engine is the Meta-Search Engine class that interfaces with all other classes to get the job done. Each class mentioned in the diagram has a unique responsibility. This makes the program highly modular in nature.

Following is a short description about each class in the Perception-based Fuzzy Meta-Search Engine:

1. **MetaSearchEngine:** This class (or rather a JSP page) takes the input keywords and a fuzzy term and processes the request by interfacing with other classes.
2. **SearchEngine:** This class queries each search engine (Google, Yahoo, AllTheWeb, AltaVista, and MSN Search) to retrieve the resultant URLs. It takes the keywords (not the fuzzy term) and passes the keywords to each search engine. Each search engine returns results in different formats. The meta-search engine removes the useless URLs (URLs that refer to advertisements) and keeps the relevant ones in a systematic way. It has some customization code to retrieve the relevant results by eliminating certain advertisements and sponsored results. There two primary methods used to query the search engines:
 - a. Using HTTP connection and parsing the search engine results in HTML (Yahoo, AltaVista, AllTheWeb, and MSN Search).
 - b. Using an API provided by the search engine (Google).
3. **HTMLPageParser:** This class takes a URL and loads the URL in memory by opening an HTTP connection and loading the HTML tags into an object using HTML

Parser. It then iterates through the collection of HTML tags to parse the web page contents such as Title, Description, Keywords, and Body tags. It returns this collected information about the URL.

4. **FuzzyMembership:** This class takes a Fuzzy term as provided by the user and searches the perception database for its perception group. If the fuzzy term exists in the database, the class creates a fuzzy set containing linguistic-to-linguistic mapping. Basically, the fuzzy set consists of other fuzzy terms in that perception group and their fuzzy memberships based on their closeness to the fuzzy term being queried. The fuzzy membership is calculated based on the formula:

$$\text{Membership} = e^{-\frac{(x - \bar{x})^2}{\sigma}} \text{-----(5.1)}$$

Where x = Score/value of one of the fuzzy terms in the perception group,

\bar{x} = Score/value of the fuzzy term being queried,

σ = Sigma (width factor of the fuzzy membership function)

5. **LingToLingRanker:** The LingToLingRanker class performs Linguistic-To-Linguistic” matching and ranking. It takes the results obtained from SearchEngine and HTMLPageParser (URLs and their contents) and returns an overall highest fuzzy relevance score for each URL. For each fuzzy term in the perception group starting with the one with the highest fuzzy membership, the LingToLingRanker class interfaces with the FuzzyRanker class and DistanceRanker class to retrieve fuzzy and distance scores. It then multiplies the scores to get an overall score for that fuzzy term. The LingToLingRanker class keeps track of the maximum overall score for a

URL among all the fuzzy terms. The fuzzy term that gets the maximum score is considered the best fuzzy match in the URL.

6. **FuzzyRanker:** The FuzzyRanker class takes a URL and its details to parse for the given fuzzy term. If the fuzzy term exists, it then queries the fuzzy vector to retrieve its membership and return that membership value as a fuzzy score.
7. **DistanceRanker:** The DistanceRanker class takes a URL and its details to deduce the distance between the given fuzzy term and the keyword. It breaks the textual contents of a URL into a vector of sentences using the SentenceLexer class. Within a sentence, if the fuzzy term is adjacent to the keyword, it gets the highest distance score. If the fuzzy term is in the same sentence as the keyword, it gets a slightly lower score. If the fuzzy term is in the same page as the keyword, then it gets yet a smaller score.
8. **SentenceLexer:** The SentenceLexer class takes a chunk of textual data and breaks it down into tokens based on a specified grammatical definition of a token. In this context, a sentence is considered as a token. Thus, the SentenceLexer class breaks the textual data into a vector of sentences. This class was generated by the JFLEX Lexer generator package. To generate this java class file, one has to give a grammar definition file to the JFLEX lexer generator application. So, a grammar definition file containing the definition of a sentence was created. A sentence is a bunch of words that end with a period (.) or a semi-colon (;) or an exclamation mark (!) or a question mark (?).

9. **LingToNumRanker:** The LingToNumRanker class performs “Linguistic-To-Numeric” matching and ranking. It takes the results obtained from SearchEngine and HTMLPageParser (URLs and their contents) and returns a fuzzy numeric score for each URL. It uses the NumericRanker class to get a fuzzy score for each web page.
10. **NumericRanker:** It takes the fuzzy term being queried and creates a fuzzy set containing linguistic-to-numeric mapping of fuzzy terms to their ideal numeric values based on pre-defined fuzzy rules. It then parses the body of the web page to get the actual value of the perception. The NumericRanker then determines the fuzzy membership of the web page based on the ideal score of the fuzzy term being queried and the actual value.

The fuzzy membership is calculated based on the formula:

$$\text{Membership} = e^{-\frac{(x - \bar{x})^2}{\sigma}} \text{-----(5.2)}$$

Where x = Actual value of the perception as found in the web page,

\bar{x} = ideal value of the perception as computed based on fuzzy rules,

σ = Sigma (width factor of the fuzzy membership function)

11. **XMLPresenter:** The XMLPresenter class retrieves the consolidated and ranked results from the Perception-based Fuzzy Meta-Search Engine and presents the results in an XML format coupled with and XML Style Sheet (XSL). It also takes user’s preferences into account by looking into the cookies collection to display fuzzy relevance scores.
12. **ErrorModule:** The ErrorModule class logs errors to a file if they occur.

13. **Preferences:** The Preference class is actually a JSP page that collects and stores user's preferences as cookies such that the XMLPresenter class can read the user's preference from the cookies. Some preferences include whether or not user would like to view the Linguistic-to-Linguistic or Linguistic-to-Numeric scores. Also, users can choose which search engines the system should query for their search.

5.2 Logic flow chart

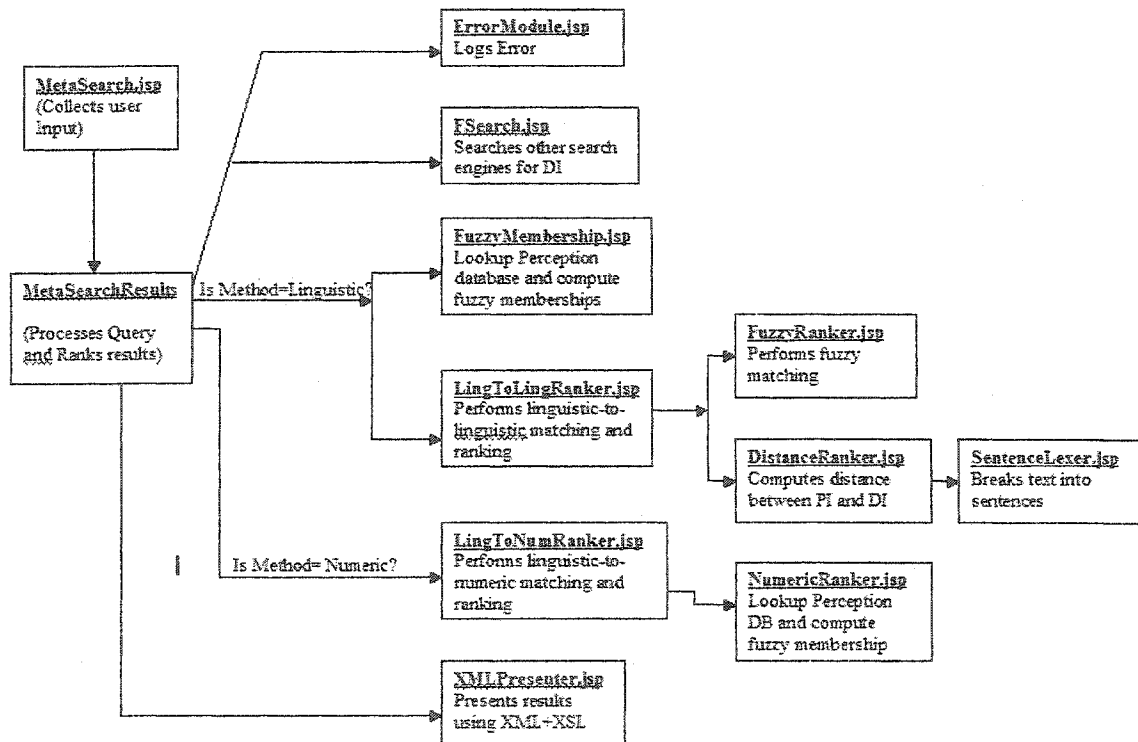


Figure 5.2 – Fuzzy Meta-Search Engine Logic Flowchart

The diagram above describes the logical flow of how different modules co-ordinate to accomplish the task of searching and presenting results. The MetaSearch.jsp is the point of entry for user to provide the input query. Then, the control is transferred to

MetaSearchResults.jsp that utilizes different classes to perform the search and present results. First, it uses the FSearch.jsp that performs search of the keyword against other search engines. Second, if the type of search is Linguistic-to-Linguistic, then it uses LingToLingRanker.jsp to filter and rank the results according to the proposed Linguistic-To-Linguistic methodology. Else, if the type of search is Linguistic-To-Numeric, then it uses LingToNumRanker.jsp to filter and rank the results according to the proposed Linguistic-To-Numeric methodology. If the type of search is standard search, then no filtering or ranking of results occurs. Finally, the MetaSearchResults.jsp uses the XMLPresenter.jsp to present the results in an XML DOM format. The browser then applies the style sheet to the XML results to present the results in a user-friendly format.

6 Evaluation of Results

The concept of using GCL to improve the search is powerful and is backed by an accepted Computational Theory of Perceptions. In this evaluation, comparison is performed on Standard Search vs. Fuzzy Linguistic-to-Linguistic Search and Fuzzy Linguistic-to-Numeric Search.

There have been two similar search tools developed as part of the project. One is the meta-search tool and another is the local/domain-specific search tool. In Meta-Search, the search is conducted by querying multiple search engines. In the local/domain-specific search tool, a collection of web sites pertaining to a specific domain has been pre-indexed by crawler software called PhpMySearch. As mentioned in Section 3.2.1, a domain of food and health has been chosen. So, a few food and health web sites have been pre-indexed using PhpMySearch. PhpMySearch crawls through the given web sites and creates a database containing information such as URL, title, description, body, etc for each HTML page that it visits. In addition, PhpMySearch also performs the search against the indexed web sites to return matching results.

6.1 Standard Search vs. Fuzzy Linguistic-to-Linguistic Search

In this section, standard search is compared against the proposed Fuzzy Linguistic-To-Linguistic search in terms of quality and quantity. 20 random fuzzy queries related to the food and health domain have been performed on a Standard and Fuzzy Linguistic-to-Linguistic search using two applications (Meta-Search and Local Search). In each method, the results of the search were visually inspected and the number of relevant

results was determined. Then, an overall percentage of relevance was determined by dividing the (total number of relevant URLs for 20 queries) by the (total number of URLs for the same 20 queries) in the result set. Then, the overall improvement percentage over the standard search results to determine the effectiveness of the proposed Linguistic-to-Linguistic search method.

$$\text{Overall Relevance Percentage} = \frac{\text{Total number of relevant URLs for 20 queries}}{\text{Total number of URLs for 20 queries}} \text{-----(6.1)}$$

$$\text{Overall Improvement over Standard Search} = \frac{\text{Overall Relevance Percentage (L2L)}}{\text{Overall Relevance Percentage (Standard)}} \text{-----(6.2)}$$

6.1.1 Local/Domain-Specific Search

When searched for 20 random queries, the results returned by the two approaches, using the Local search application, are summarized in Table 6.1. Each query contains a fuzzy and a crisp word. For instance, in “prevent cancer” the word “prevent” is the fuzzy word and “cancer” is the crisp word. Table 6.1 proves that quantitatively, the Linguistic-To-Linguistic search method returned fewer and more relevant URLs. The proposed Linguistic-To-Linguistic method shows 66% improvement in the search results. Not only that, the fuzzy Linguistic-To-Linguistic search approach shows the results from the most relevance to the least relevance. This proves the filtering and ranking capabilities of the fuzzy search methodology.

Local/Domain-Specific Search						
Keyword	Standard Search			Fuzzy Linguistic-To-Linguistic Search		
	No. of URLs	No. of Relevant URLs	Relevance %	No. of URLs	No. of Relevant URLs	Relevance %
prevent cancer	228	134	58.77	128	115	89.84
reduce cholesterol	275	222	80.73	213	213	100.00
reduce diabetes	197	60	30.46	93	45	48.39
Increase energy	119	71	59.66	66	51	77.27
relieve diarrhea	64	31	48.44	26	25	96.15
relieve constipation	73	44	60.27	39	39	100.00
Treat fever	42	11	26.19	9	8	88.89
Treat migraine	59	28	47.46	24	24	100.00
reduce menstrual cramps	147	11	7.48	14	10	71.43
Treat ulcer	40	10	25.00	9	9	100.00
beneficial honey	131	75	57.25	64	64	100.00
significant calcium	138	102	73.91	91	89	97.80
significant sodium	196	159	81.12	133	116	87.22
rich folate	194	82	42.27	76	76	100.00
rich protein	266	205	77.07	165	153	92.73
Improve vision	93	11	11.83	10	10	100.00
Cure depression	67	29	43.28	24	24	100.00
muscular bones	35	25	71.43	26	23	88.46
improve memory	95	15	15.79	11	11	100.00
Treat asthma	125	62	49.60	90	62	68.89
OVERALL RELEVANCE						
	2584	1387	54%	1311	1167	89%
IMPROVEMENT PERCENTANGE						166%

Table 6.1 – Standard vs. Fuzzy L2L Search using Local/Domain-Specific Search

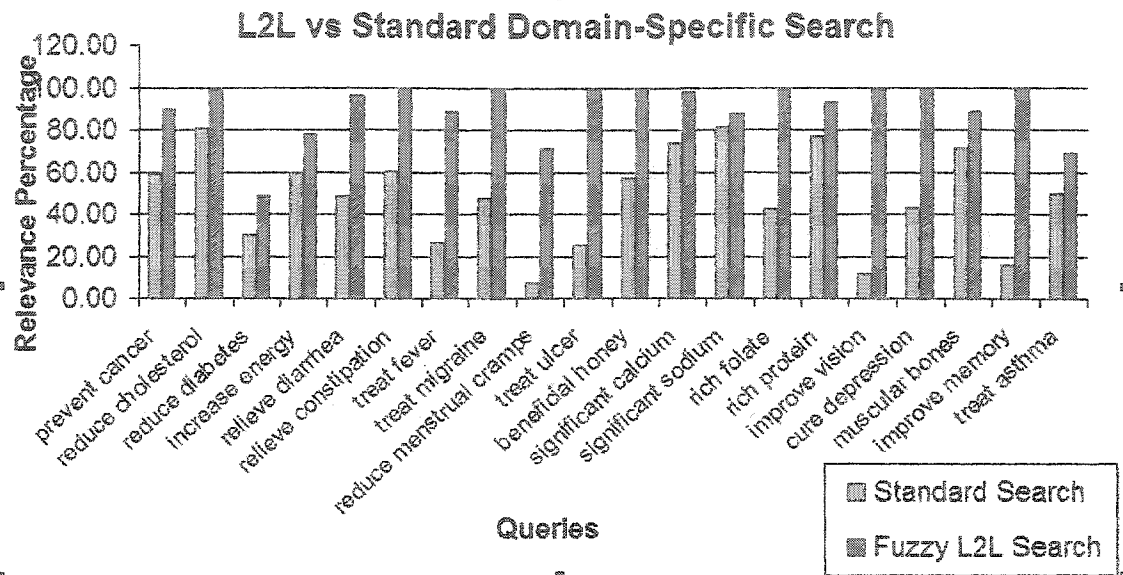


Figure 6.1 – Graphical Representation of Standard vs. L2L Search using Local/Domain-Specific Search

6.1.2 Meta-Search

When searched for the same 20 queries, the results returned by the two approaches, using the Meta-search application, are summarized in Table 6.2. Table 6.2 proves that quantitatively, the Linguistic-To-Linguistic search method returned fewer and more relevant URLs. The proposed Linguistic-To-Linguistic method shows 8% improvement in the search results. The improvement percent in this case is not very impressive because of one main reason: the first 10 results from each search engine (google, yahoo, etc.), that queries billions of web pages, are an exact match to the query even though the query has two words (fuzzy and crisp). This makes the standard search pretty powerful. However, if one visually inspects the results, it will be quickly evident that the fuzzy Linguistic-To-Linguistic search approach improves upon the ranking of the results

provided by the standard search thus providing the user with the most relevant results at the top.

Keyword	Meta-Search					
	Standard Meta-Search			Fuzzy Linguistic-To-Linguistic Meta-Search		
	No. of URLs	No. of Relevant URLs	Relevance %	No. of URLs	No. of Relevant URLs	Relevance %
prevent cancer	31	30	96.77	15	15	100.00
reduce cholesterol	31	26	83.87	33	33	100.00
reduce diabetes	40	38	95.00	11	11	100.00
Increase energy	33	27	81.82	10	10	100.00
relieve diarrhea	35	32	91.43	22	22	100.00
relieve constipation	37	36	97.30	27	27	100.00
treat fever	39	39	100.00	19	19	100.00
treat migraine	27	27	100.00	16	16	100.00
reduce menstrual cramps	35	34	97.14	25	25	100.00
treat ulcer	33	33	100.00	20	19	95.00
beneficial honey	27	27	100.00	15	15	100.00
significant calcium	38	34	89.47	22	19	86.36
significant sodium	30	13	43.33	15	15	100.00
rich folate	34	34	100.00	24	22	91.67
rich protein	42	39	92.86	3	3	100.00
Improve vision	39	29	74.36	13	12	92.31
Cure depression	38	32	84.21	11	11	100.00
muscular bones	29	29	100.00	8	7	87.50
Improve memory	41	34	82.93	17	17	100.00
treat asthma	33	33	100.00	18	18	100.00
OVERALL RELEVANCE						
	692	626	90%	344	336	98%
IMPROVEMENT PERCENTANGE						108%

Table 6.2 – Standard vs. Fuzzy L2L Search using Meta-Search

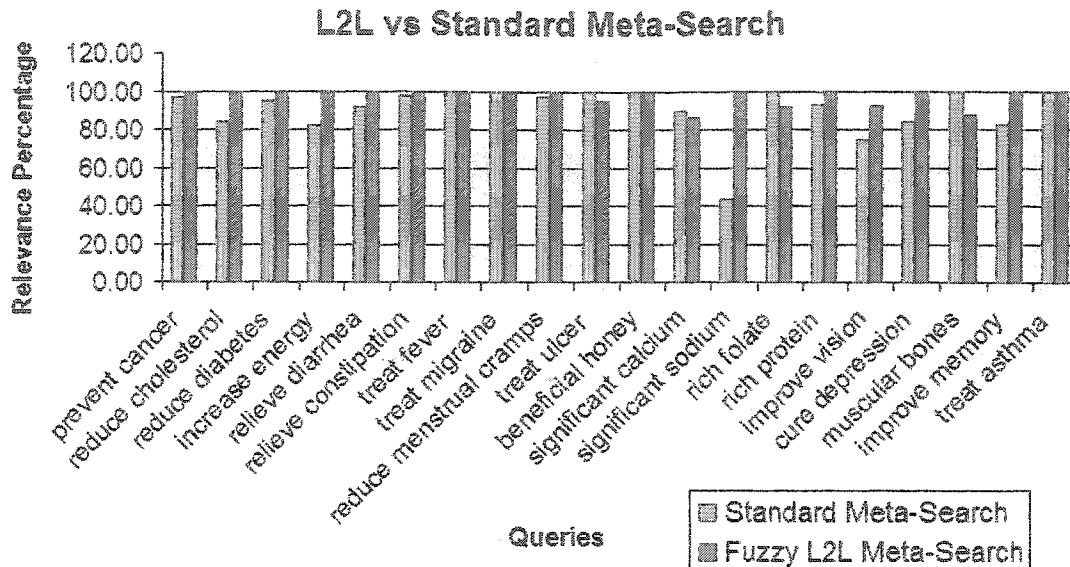


Figure 6.2 – Graphical Representation of the Standard vs. L2L using Meta-Search

6.2 Standard Search vs. Fuzzy Linguistic-to-Numeric Search

In this section, Standard search is compared against the proposed Fuzzy Linguistic-To-Numeric search in terms of quality and quantity. As described in previous sections, in the Linguistic-to-Numeric Search, a fuzzy linguistic variable is mapped to a numeric value. In this project, the Linguistic-To-Numeric search has been limited to the subject of resourcefulness of various vitamins and minerals in the domain of food and health. Also, there are only three possible fuzzy terms for each crisp term: high, medium, and low. 20 random fuzzy queries related to the food and health domain have been performed on a Standard and Fuzzy Linguistic-to-Numeric search using two applications (Meta-Search and Local Search). In each method, the results of the search were visually inspected and the number of relevant results was determined. Then, an overall

percentage of relevance was determined by dividing the total number of relevant URLs for 20 queries by the total number of URLs for the same 20 queries in the result set. Then, the overall improvement percentage over the standard search results to determine the effectiveness of the proposed Linguistic-to-Numeric search method.

$$\text{Overall Relevance Percentage} = \frac{\text{Total number of relevant URLs for 20 queries}}{\text{Total number of URLs for 20 queries}} \text{----- (6.3)}$$

$$\text{Overall Improvement over Standard Search} = \frac{\text{Overall Relevance Percentage (L2N)}}{\text{Overall Relevance Percentage (Standard)}} \text{----- (6.4)}$$

6.2.1 Local/Domain-Specific Search

When queried for 20 random queries, the results returned by the two approaches, using the Local search application, are summarized in Table 6.3. Each query contains a fuzzy and a crisp word. For instance, in “high calcium” the word “high” is the fuzzy word and “calcium” is the crisp word. Table 6.3 proves that quantitatively, the Linguistic-To-Numeric search method returned fewer and more relevant URLs. The proposed Linguistic-To- Numeric method shows 23% improvement in the search results. Not only that, the fuzzy Linguistic-To- Numeric search approach shows the results from the most relevant to the least relevant.

Local/Domain-Specific Search						
Keyword	Standard Search			Fuzzy Linguistic-To-Numeric Search		
	No. of URLs	No. of Relevant URLs	Relevance %	No. of URLs	No. of Relevant URLs	Relevance %
high calcium	224	224	100.00	145	145	100.00
low calories	233	233	100.00	129	129	100.00
high folate	202	201	99.50	167	166	99.40
high iron	212	212	100.00	156	156	100.00
medium magnesium	260	220	84.62	172	172	100.00
high potassium	233	233	100.00	176	176	100.00
low saturated fat	267	267	100.00	118	118	100.00
low sodium	274	274	100.00	251	251	100.00
high protein	134	134	100.00	276	276	100.00
high vitamin+a	392	130	33.16	156	135	86.54
high vitamin+c	449	248	55.23	207	177	85.51
medium vitamin+e	70	21	30.00	138	138	100.00
medium vitamin+k	70	4	5.71	8	8	100.00
low carbohydrate	262	262	100.00	245	245	100.00
low cholesterol	517	369	71.37	254	254	100.00
high fat	496	369	74.40	266	266	100.00
high saturated fat	267	267	100.00	120	120	100.00
high fiber	466	389	83.48	295	259	87.80
high biotin	14	14	100.00	6	6	100.00
high vitamin+d	136	17	12.50	16	16	100.00
OVERALL RELEVANCE						
	5178	4088	79%	3301	3213	97%
IMPROVEMENT PERCENTAGE						123%

Table 6.3 – Standard vs. Fuzzy L2N Search using Local/Domain-Specific Search

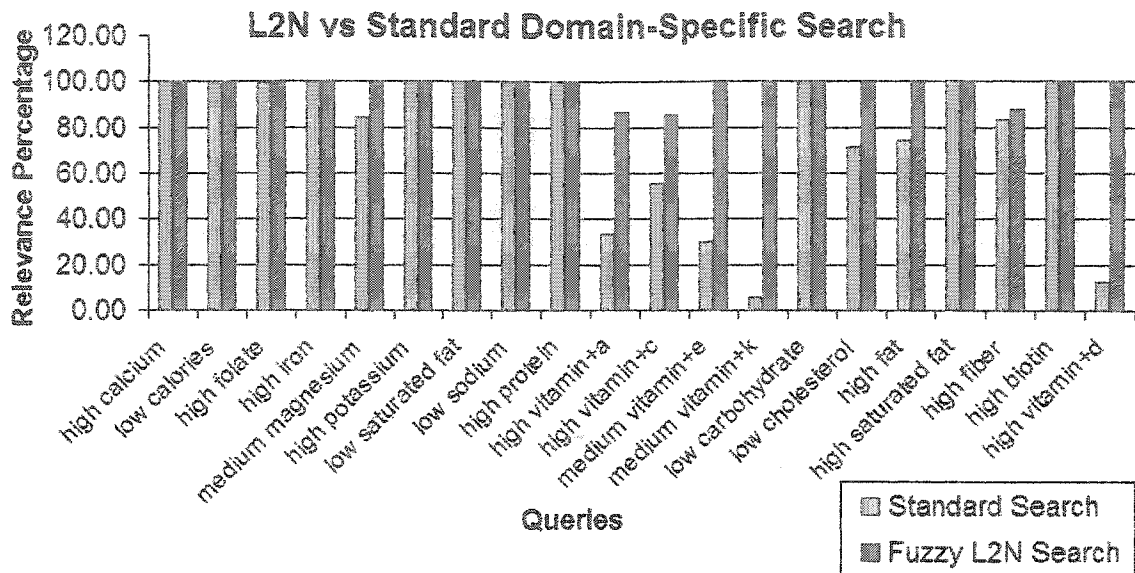


Figure 6.3 – Graphical Representation of Standard vs. L2N using Local/Domain-Specific Search

6.2.2 Meta-Search

When queried for the same 20 queries using the Meta-Search application, the results returned by the two approaches are summarized in Table 6.4. Table 6.4 proves that quantitatively, the Linguistic-To-Numeric search method returned fewer and more relevant URLs. The proposed Linguistic-To-Numeric method shows 11% improvement in the search results. The improvement percentage in this case is not so impressive because the Linguistic-To-Numeric search depends upon availability of a numeric value in the vicinity of the keyword such that the numeric value is used to perform fuzzy ranking. However, most web pages do not have the data in the required format leading to very few hits and even fewer relevant URLs.

Meta-Search						
Keyword	Standard Meta-Search			Fuzzy Linguistic-To-Numeric Meta-Search		
	No. of URLs	No. of Relevant URLs	Relevance %	No. of URLs	No. of Relevant URLs	Relevance %
high calcium	37	34	91.89	0	0	0.00
low calories	39	34	87.18	1	1	100.00
high folate	34	34	100.00	4	4	100.00
high iron	36	34	94.44	3	3	100.00
medium magnesium	34	34	100.00	1	1	100.00
high potassium	40	39	97.50	2	2	100.00
low saturated fat	34	28	82.35	4	4	100.00
low sodium	31	26	83.87	2	2	100.00
high protein	40	39	97.50	5	5	100.00
high vitamin+a	46	32	69.57	1	1	100.00
high vitamin+c	48	41	85.42	2	2	100.00
medium vitamin+e	44	37	84.09	1	1	100.00
medium vitamin+k	38	30	78.95	4	4	100.00
low carbohydrate	35	32	91.43	1	1	100.00
low cholesterol	46	42	91.30	5	5	100.00
high fat	38	38	100.00	4	4	100.00
high saturated fat	35	34	97.14	6	6	100.00
high fiber	29	28	96.55	4	4	100.00
high biotin	34	33	97.06	1	1	100.00
high vitamin+d	46	37	80.43	2	2	100.00
OVERALL RELEVANCE						
	764	686	90%	53	53	100%
IMPROVEMENT PERCENTANGE						111%

Table 6.4 – Standard vs. Fuzzy L2N Search using Meta-Search

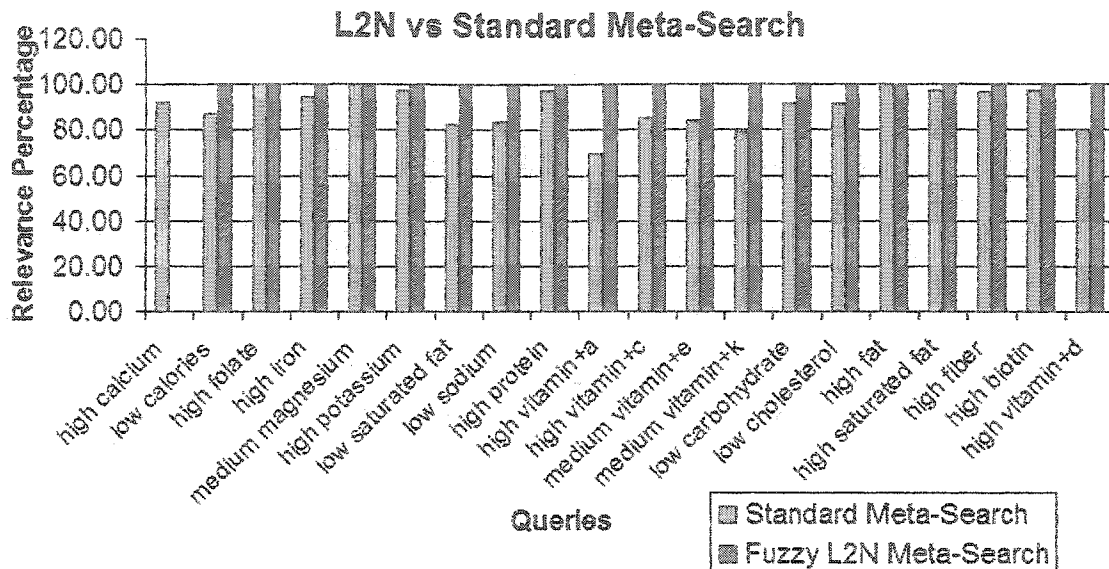


Figure 6.4 – Graphical Representation of Standard vs. L2N using Meta-Search

In conclusion,

1. The Linguistic-To-Linguistic method has proven to be most effective since it performs good filtering for relevant URLs by using linguistic variables.
2. The local search shows much better improvement than meta-search in both the methods of Linguistic-To-Linguistic and Linguistic-To-Numeric searches. This, once again, proves that the proposed method works best on a very specific domain instead of the entire World Wide Web. In addition, the specific domain that contains imprecise data/information will benefit from this method.
3. Given a specific domain containing data in a pre-specified format, even the Linguistic-To-Numeric search displays the most relevant results.
4. The use of fuzzy linguistic variables allows for better *filtering* of relevant results and the use of fuzzy logic allows for better *ranking* the relevant URLs. This allows for a more reliable and user-friendly search system.

7 Conclusion

The conventional search methods existing today lack the power of deductive capabilities. By studying a small portion of human behavior during the search process, certain concepts and methodologies can be modeled to improve the search systems dramatically. Given the imprecise and unstructured web content, the usage of perception database combined with fuzzy logic has resulted in language flexibility and relevance as features of the Perception-based Fuzzy Meta-Search Engine. By allowing the users to restrict the keywords using a fuzzy term (perception-based) in Linguistic-To-Linguistic search, the meta-search engine provides enormous expressive power by returning results that are relevant to the user's query. By mapping linguistic variables to numeric values in the Linguistic-To-Numeric search, the search interface has become more human-friendly. Combining the perception-based search with a meta-search that queries many search engines at once creates a fine information retrieval system. Having a meta-search engine gives the users the best of many worlds. Also, the fact that the results are advertisement-free gives the users a chance to focus on their task at hand. The Perception-based Fuzzy Meta-Search Engine looks and behaves like any standard search engine yet produces more relevant results; this fact makes it very powerful and reliable. However, it must be noted that such deductive capabilities are currently limited to one domain because the fuzzy variables may have different weights in different domains. This project can be considered as one step forward in implementing an intelligent search engine using fuzzy logic to yield a small, precise, and effective result set.

References

- Choi, D.Y. (2003). Enhancing the power of Web search engines by means of fuzzy query. *Decision Support Systems*, 35, 31-44.
- Kantrowitz, M. (1993). What is fuzzy logic? *Computer Science Department, Carnegie Mellon University*. Retrieved February 7, 2003, from the World Wide Web: <http://www-2.cs.cmu.edu/Groups/AI/html/faqs/ai/fuzzy/part1/faq-doc-2.html>
- McCool, R. & Guha, R.V., (2002). TAP: A System for integrating Web Services into a Global Knowledge Base. Retrieved April 4, 2003, from the World Wide Web: <http://tap.stanford.edu/tap/papers.html>
- Nikravesh, M. (2002). Computing with Words, Computational Theory of Perceptions, Precisiated Natural Language, and Perception-Based Decision Analysis. Berkeley Initiative in Soft Computing. Abstract retrieved February 7, 2003, from the World Wide Web: <http://buffy.eecs.berkeley.edu/ResearchSummary/02abstracts/nikravesh.4.html>
- Nikravesh, M. (2002). The BISC Initiative: Fuzzy Logic and the Internet (FLINT); Perception Based Information Processing and Analysis. Berkeley Initiative in Soft Computing. Abstract retrieved February 7, 2003, from the World Wide Web: <http://buffy.eecs.berkeley.edu/ResearchSummary/02abstracts/nikravesh.7.html>
- Sullivan, D. (2002). How Search Engines Work. *Search Engine Watch*. Retrieved February 7, 2003, from the World Wide Web: <http://searchenginewatch.com/webmasters/work.html>
- Teo, D. (1995). Fuzzy Logic and its applications. *Computer Science Department, Santa Clara University*.
- Zadeh, L. (1999). From Computing with Numbers to Computing with Words - From Manipulation of Measurements to Manipulation of Perceptions. *IEEE Transactions on Circuits and Systems*, 46(1), 105-119.