Singaporean Journal of Scientific Research

International Journal of Engineering (IJE)
Singaporean Journal of Scientific Research(SJSR)
Vol.6.No.4.2014 Pp.219-223
Available at:www.iaaet.org/sjsr
Paper Received:19-02-2014
Paper Accepted:10-03-2014
Paper Reviewed by:1Prof.Dr.Karthikeyan2. Chai ChengYue

SEARCH ENGINE OPTIMIZATION BY FUZZY CLASSIFICATION AND PREDICTION

S. Milton Ganesh
Department of Computer Science & Engineering,
University College of Engineering,
Ramanathapuram, Ramanathapuram – 623 513.
softengineermilton@gmail.com

ABSTRACT

Search Engine Optimization will last as long as there are search engines and the Internet. New techniques are being developed everyday which provide some satisfactory results than the previous ones and hence are more preferred. One such technique is proposed in this paper which takes in to account the characteristics such as page rank as usual, mouse movements of a user and the eye movements of a user while surfing a web page. The technique is proposed so that the search engine gets human like thinking and accordingly produces the search results. The search engines which are available in the market today take into account many characteristics on-page and off-page to predict the search results. But none of them take into account all the heuristic details a user is currently submitting to a search engine such a mouse movements and eye movements. The proposed algorithm takes each and every parameter in user point of view and also the usual page ranking and applies fuzzy logic intelligence to predict the subsequent search results.

Keywords: Fuzzy Logic, Mouse Movements, Eye Movements, Page Ranking, rational search

1. INTRODUCTION

Usually, a search engine consists of crawler, indexer, and ranker. A crawler retrieves web documents from the web [B. Pinkerton]. Search engines create a map of the web by indexing web pages according to keywords. The database of search engine ideally returns a list of relevant URL's, corresponding to the search keywords.

At first glance, the service search-engines seems very useful and faultless, but by a more careful examination one may notice weaknesses in their search results. The reason behind was, there are many problems in the way of search-engines and challenges come across their performances. Among them are the high volume of pages and bandwidth limitations. Also because many web pages are in the dark internet, spider is unable to find them (Gaston L'Huillier, 2011). Another cause of weakness in the search engine results can be because of their databases are not up to date. In addition to these challenge, high expectations has made the job of search engine crawlers harder. For example the crawler is expected to look for and retrieve the pages containing the synonyms of the searched keyword too, there occurs a error. Since the two strings won't be similar, all these makes the job of a crawler even harder. Nevertheless, because each language has its own exceptions and search engine are based on string processing, ideal results cannot be expected. Pages with an informal tone and slang or misspelled words will make the search-engines results more difficult. The inability of search-engines in processing parts of web pages like frames, Adobe Flash, pictures and also JavaScript and AJAX codes is an another challenge for search engines (Ricardo Baeza-Yates, 2011). In addition to normal challenges mentioned, competition over site rankings has caused some webmasters to design their websites with not so useful pages. These misconducts include:

- Repeating words in order to increase the density of the keyword searched.
- Exchanging unrelated links with websites that have a high ranking in the search engines.
- Generating dynamic pages with the purpose of deceiving search engines.

Concept-based IR – Concept-based IR is the search for information based on its meaning rather than the keywords searched [Hele-Mai Haav, Tanel-Lauri Lubi]. This approach promises to increase the quality of responses since it captures the semantics of the documents.

The biggest problem with search-engines is the use of mechanical algorithms. This paper advises search-engines to hand the job of decision making about the content of web sites to users, because humans are very much faster and have a very lower rate of error and can decide about the usefulness of a website with more justice. Also, search engines have been built to aid humans only. In this paper, an attempt has been made for search-engines which will maintain user feedback such as mouse movements, eye movements instead of page analysis alone.

2. FUZZY SETS AND FUZZY LOGIC

According to the Oxford American Dictionary, fuzzy means blurred or indistinct. Now, the last thing, searchers want are blurred or indistinct results, yet paradoxically, a liberal dose of fuzzy logic can actually improve the precision of search engine results. Using fuzzy logic, a search engine can relax the boundaries between word meanings to a certain degree [Chris Sherman]. Just as a camera lens set to a larger aperture brings a greater range of view into focus, fuzzy logic in a search engine will expand the depth of field of potential search results. Consider some human expressions like "very flexible", "easily integrated" and "good solution". This type of vague expressions are characteristic of the way we humans communicate through language and as such is an integral part of our thinking process. This contrasts sharply with the traditional Boolean logic of computer programming which deals with either true (1) or false (0) and nothing in-between. Fuzzy logic bridges the gap between them, providing a framework that allows you to numerically encode linguistic expressions and through that gives you a flexible rule-based system.

Essentially, a fuzzy set is a set whose members of the set may have degrees of membership between 0 and 1, as opposed to classical sets where each element must have either 0 or 1 as the membership degree—if 0, the element is completely outside the set; if 1, the element is completely in the set. As classical logic is based on classical set theory, fuzzy logic is based on fuzzy set theory. To be able to express ourselves with this new fuzzy thing, we need some basic rules. Our ultimate goal is to be able to define logical expressions that we later can turn into statements.

2.1 Linguistic Variables

A numerical variable takes numerical value as Age = 65. A linguistic variable takes linguistic value as Age = old. A linguistic value mentions a fuzzy set. T(age) = {young, not young, very young, …, middle aged, not middle aged, ……, old, not old, very old, … more or less old….. etc}. Each of the value for example young, not young is interpreted in terms of a membership function in fuzzy logic. Membership functions can be defined as triangular, trapezoidal, Gaussian and many other types.

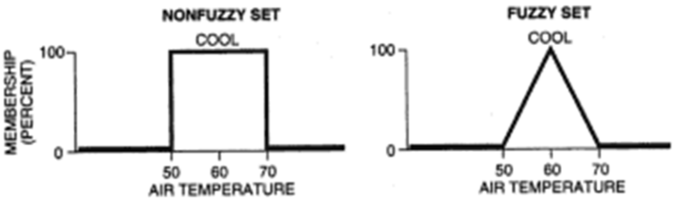


Figure1 Temperature expressed using Boolean and fuzzy logic.

2.2 Fuzzy If-Then Rules

Fuzzy is a rule-based system consists of *if-then rules*, a bunch of *facts*, and an *interpreter* controlling the application of the rules, given the facts. These *if-then* rule statements are used to formulate the conditional statements that comprise the complete knowledge base. A single *if-then* rule assumes the form ‘if *x* is *A* then *y* is *B*’ and the if-part of the rule ‘*x* is *A*’ is called the *antecedent* or *premise*, while the then-part of the rule ‘*y* is *B*’ is called the *consequent* or *conclusion*. Some fuzzy rules are mentioned below.

Rule 1: *If A and C then Y*
Rule 2: *If A and X then Z*
Rule 3: *If B then X*
Rule 4: *If Z then D*

2.3 Fuzzy Operators

There are three relevant operators in the fuzzy set logic: OR, AND and NOT. They also obviously exist for regular Boolean logic, but we need to expand their definition to support our new non-sharp membership functions:

$$\text{OR: } A \cup B = MAX(A,B)$$
$$\text{AND: } A \cap B = MIN(A,B)$$
$$\text{NOT: } \neg A = 1 - A$$

2.4 A fuzzy Engine

A fuzzy inference engine first fuzzifies the crisp input. The inference engine apply the rules on the fuzzy input and produce the fuzzy outputs. The results are converted to the crisp form before submitted to the user.

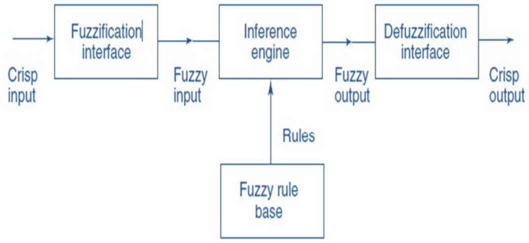
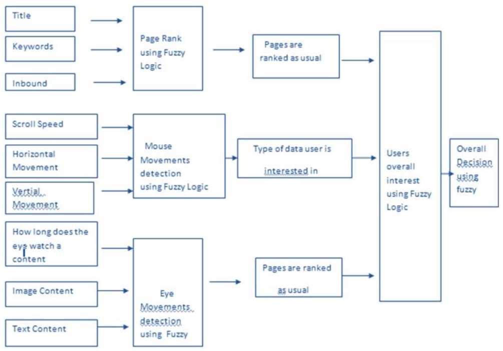


Figure2 Fuzzy Inference Engine

3. METHODOLOGY

The proposed methodology aims to provide results which should reflect the human thinking. For this purpose, the input parameters from the humans that is the users of the web are taken into consideration and the search is improved upon by these parameters as the primary ones along with the usual page ranking. The system is divided into the following steps: (1) Page Ranking (2) Mouse Movements consideration (3) Eye Movements consideration (4) Input parameters mapping and decision making (5) Fuzzy Rules for the Search Decision Making System.



Title
Keywords
Inbound
Page Rank using Fuzzy Logic
Pages are ranked as usual
Scroll Speed
Horizontal Movement
Vertial Movement
Mouse Movements detection using Fuzzy Logic
Type of data user is interested in
How long does the eye watch a content
Image Content
Text Content
Eye Movements detection using Fuzzy
Pages are ranked as usual
Users overall interest using Fuzzy Logic
Overall Decision using fuzzy

Figure3 Hybrid Fuzzy Search Engine with user input and page ranking methods.

3.1 Page Ranking

PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. If the number inbound links to a web page is more, then it is considered very important. Page ranking is also done based on a number of other factors as well. Most important of factors such as title, keywords, inbound links are considered in this paper. Each one of the page ranking parameters considered are expressed in terms of linguistic variables and the associated linguistic values. A separate fuzzy inference system is employed to find out the page rank taking into consideration the parameters as title, keywords, number of inbound links.

For example, the parameter title is a linguistic variable and its linguistic values are low, medium, high. Accordingly, for each of the other two linguistic variables, linguistic values are assigned.

Table1. Fuzzy sets for the input metrics

Search Engine Parameter	Input Metrics	Linguistic Label	Membership Degree
PageRank	Title	Low, Medium, high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}
	Keyword	Very low, medium, high, very high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}, {0,0.5,1}
	Inbound Links	Low, Medium, high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}
Mouse Movements	Scroll Speed	Low, Medium, high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}
	Horizontal Movements	Very low, medium, high, very high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}, {0,0.5,1}
	Vertical Movements	Low, Medium, high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}
Eye Movements	The data user is interested actually interested in	Low, Medium, high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}

	Text Content	Low, Medium, high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}
	Image Content	Low, Medium, high	{0,0.5,1}, {0,0.5,1}, {0,0.5,1}

3.2 Mouse Movements Consideration

To survey the user behavior while dealing with undesired pages, a webpage may be designed and the statistical society can be asked to find the most interesting philosophical passage within that page. Of course, the page should not just filled with philosophical passages and should contain four parts: The first part of the page includes a number of image ads that had no relevance to the topic that the user was looking for and it is anticipated that the user would skip this part very fast. In the second part, a hamburger receipt should be described. Obviously this part has nothing to do with what the users will be looking for, either. As the part before, it is expected that the users would also skim this part rather quickly, but since the topic of this part is not recognizable as easily as the picture ads, it is expected that users would go through it with a somewhat lesser speed. The third part could be filled with different passages but user shouldn't be able to find what he/she is looking for in this part either because these passages contained facts and figures while users where looking for the most interesting philosophical passage. Last part was where users are expected to spend more time and go through with more attention, and it may contain may philosophical passages. The mentioned web page should contain a JavaScript code that, based on user behavior, captured vertical mouse speed, scroll speed, horizontal mouse speed, duration time and standard deviation of horizontal mouse movement and passes them to a database. The parameters listed for each four sections of the site need to be separately calculated and recorded in the database. A total of atleast 50 computer students can participated in this experiment. As mentioned, participants can be asked to choose the most interesting passage within the page. Of course users will not be aware that their movements and behavior were recorded at the time. In other words, users will think that they are participating in a survey, because if they knew their movements were being recorded, it could affect their natural behavior.

A separate fuzzy inference system is employed considering the mouse parameters mouse speed, horizontal movements and vertical movements and produces a fuzzy output. From the output of the system, we can accurately predict the user behavior with respect to the mouse movements.

3.3 Eye Movements Consideration

As separate fuzzy inference system is employed for finding the eye movements of a human who surfs the web pages. Using a web camera, the movements of eyes such as horizontal movements, vertical movements, how long a user sees a web page, whether the content viewed is a text content or the content is an image content. Eye movements are very important that they actually reflect the users' view of the web page. When a user surfs a web page for some content, he might be interested in images or he could be interested in text-content. It is opt to find out where the interest of a user of a web page lies in. If the user prefers to see images and spends more time in looking at images to get the information for which he is surfing the net, then the next search could focus on bring out searches with more image based contents. On the other hand, if the user prefers to look at the text based content rather than image based ones, then the next search could focus on brining out pages with more text contents. This way, we can reduce the download time for a user surfing the net as it is evident that downloading a page with more images takes more time than a page with only text content. A fuzzy inference system is employed to find out the eye characteristics of a user and the output of this inference system is supplied as input to the final decision making system.

3.4 Parameters Mapping and Decision Making

If the search results are based only on the page ranking techniques as used in most of the web browsers in the market today, then the users' perspective and rationality of search engine is lost. On the other hand, if the searching is based only on the users interaction with the web pages and not on the on-page and off-page contents in the web pages, then the actual concept of a search engine is defeated. Thus, it is better if we consider the both of the parameters. Thus the fuzzified output of Page Ranking method, fuzzified output of mouse movements consideration and the fuzzified output of eye movement consideration are considered to a single fuzzy system and produces a single output combining all the afore mentioned parameters. When the user first searches, the page ranking technique alone is employed. But as the user progresses with three or more pages, then the proposed algorithm works at its best and produces optimal result which would surely enhance the user's expectation of a search engine.

3.5 Fuzzy Rules

3.5.1 Fuzzy Rules for PageRanking

* If (Inbound Links is high) or (title is low) then PageRank is high

* If(Inbound Links is low) or (title is high) then PageRank is medium

* If(title is high) and (keyword is high) then PageRank is high
* If (title is low) and (keyword is low) then PageRank is low
* If (title is medium) and (keyword is low) then PageRank is low

3.5.2 Fuzzy Rules for Mouse Movements

- IF (ScrollSpeed is VeryHigh) OR (VerticalMouseSpeed is VeryHigh) OR (StandardDeviation is VeryLow) OR (HorizontalMouseSpeed is VeryHigh) OR (DurationTime is VeryLow) THEN (MouseMovements is VeryLow)
- IF (ScrollSpeed is High) OR (VerticalMouseSpeed is High) OR (StandardDeviation is Low) OR (HorizontalMouseSpeed is High) OR (DurationTime is Low) THEN (MouseMovements is VeryLow)
- IF (ScrollSpeed is VeryLow) OR (VerticalMouseSpeed is VeryLow) OR (StandardDeviation is High) OR (HorizontalMouseSpeed is VeryLow) OR (DurationTime is High) THEN (MouseMovements is High)

3.5.3 Fuzzy Rules for Eye Movements

* If (waiting time is high) AND (TextContent is high) THEN TextInterest is high
* If(waiting time is high) AND (ImageContent is high) THEN ImageInterest is high
* If (waiting time is low) AND (TextContent is high) THEN ImageInterest is high
* If(waiting time is low) AND (ImageContent is high) THEN TextInterest is high

3.5.4 Fuzzy Rules for the Search Decision Making System

- If (PageRanking is low) AND (MouseMovements is high) AND (TextInterest is high) THEN Interest is high
- If (PageRanking is low) AND (MouseMovements is low) AND (TextInterest is high) THEN TextInterest is Medium
- If (PageRanking is high) AND (MouseMovements is medium) AND (TextInterest is high) THEN Interest is high
- If (PageRanking is low) AND (MouseMovements is Medium) AND (ImageInterest is Medium) THEN Interest is Medium
- If (PageRanking is Medium) AND (MouseMovements is high) AND (ImageInterest is low) THEN Interest is Medium.

4. CONCLUSION

The proposed search engine optimization technique can be implemented taking into account the characteristics such as page rank as usual, mouse movements of a user and the eye movements of a user while surfing a web page. Because of the right mix of the user characteristics and web page characteristics, the search engine will be able to out-perform the existing ones. The proposed technique can be tested amongst a group of at least 100 people to test its accuracy and the in future more user characteristics can be considered to further improve the accuracy of the fuzzy inference system.

REFERENCES

1. B. Pinkerton, "Finding what people want: Experiences with the webcrawler," The Second International WWW Conference, Chicago, USA, 1994
2. Gaston L'Huillier, Hector Alvarez, Sebastián A. Ríos, Felipe Aguilera (2011). Topic-based social network analysis for virtual communities of interests in the dark web. s.l. : ACM *SIGKDD Explorations Newsletter*.
3. Ricardo Baeza-Yates, Andrei Z. Broder, Yoelle Maarek (2011). The new frontier of web search technology: seven challenges. s.l. : *Search computing.*
4. Hele-Mai Haav, Tanel-Lauri Lubi, "A Survey of Concept-based Information Retrieval Tools on the Web"
5. Chris Sherman, "In Praise of Fuzzy Searching", June 6, 2001.
6. Jeff Huang, Ryen W. White, Susan Dumais (2011). No clicks, no problem: using cursor movements to understand and improve search. s.l. : CHI '11 *Proceedings of the 2011 annual conference on Human factors in computing systems.*
7. 10. Cooke, Lynne (2006). Is the Mouse a "Poor man's Eye Tracker"? s.l. : Usability and Information Design.