Proc. of the Intl. Conf. on Advances in Image Processing and Computation Techniques - PCT 2012

ISBN: 978-981-07-2969-1

Blogs Search Engine Using RSS Syndication and Fuzzy Parameters

Athraa Jasim Mohammed
School of Computing
College of Arts and Sciences
Universiti Utara Malaysia
Sintok, Kedah, Malaysia

Husniza Husni
School of Computing
College of Arts and Sciences
Universiti Utara Malaysia
Sintok, Kedah, Malaysia

Abstract— **The rapid development of the internet eventually increases the number of internet users triggering the need for an intelligent search engine that is able to minimize the search on world wide web (WWW) and find relevant information as requested. To overcome the issue of finding relevant information as well as minimizing the search on WWW, this paper proposes a search engine that is specifically designed and built using RSS syndication and fuzzy Parameters to search for information contained in blogs. The blogs search engine consists of three main phases: 1) crawling using RSS feeds algorithm; 2) indexing weblogs algorithm; and 3) searching technique using fuzzy logic. In RSS crawling process, the RSS feeds need to be gathered to extract useful information such as title, links, time published, and description. Next, indexing weblogs uses the links to retrieve the blog sites for text processing and for constructing the indexing database. In order to retrieve such information requested or queried by any user, an interface is provided to enable the blog search based on keyword with associated degree of importance. The density of keyword is then computed from the indexing database. The rank of the pages is computed by using fuzzy weighted average. The experiment resulted in mean average precision of 81.7% of total system performance.**

Keywords—Rss feeds, blog ssearch engine, fuzzy weighted average, keyword density.

I. INTRODUCTION

In the last few years, the large population of Internet communities have caused massive quantities of web data, which led to the development and consumption of information. This situation also increases and somehow motivates more people to use blogs. A weblog or blog is a "frequently updated Web page with dated entries in reverse chronological order, usually containing links with commentary" [1]. Instead of encountering blogs by chance during navigating the Internet, a search engine becomes necessary to be able to actively find interested blogs [2]. A search engine is simply "a web site used to easily locate internet resources". Search engines have facilitated the information retrieval process by adopting techniques such as Artificial Intelligence [3]. In this paper, we proposed to use two techniques – RSS technique in crawling phase and fuzzy logic in search phase. RSS (Really Simple Syndication, a web content syndicate format) is being used extensively to describe the content and related information of weblogs and news sites. The RSS data published by websites as abstract of its latest contents [4]. RSS is an XML file format designed for web content syndication whereas fuzzy logic employs Fuzzy weighted average (FWA) calculated by two parameters: the importance of query and query density. Figure 1 depicts an example of RSS feed extracted from a weblog site, which exemplifies the main parts of a RSS feed.

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
<channel>
        <title>Robotics Zeitgeist</title>
        <link>http://robotzeitgeist.com</link>
        <description>Artificial Intelligence and Robotics
      blog</description>
        <lastBuildDate>Tue, 23 Aug 2011 05:22:19
+0000</lastBuildDate>
  <item>
        <title>MABEL two-legged robot fastest in the world</title>
        <link>http://feedproxy.google.com/~r/ArtificialIntelligence
                AndRobotics/~3/5u4C_ssd790/mabel-two-legged-
                robot-fastest-in-the-world.html</link>
        <pubDate>Tue, 23 Aug 2011 05:22:19 +0000</pubDate>
        <description>It would appear that we have a new champion
                in the &#8220;what robot can run
                fastest race&#8221;. The two-legged
                robot MABEL under development for
                several years at the University of
                Michigan was recently revealed to
                reach a top running speed of 6.8 miles
                per hour or roughly 11 kilometers per
                hour. This means that MABEL is
                significantly faster than the previous
                record holder which was
                Toyota&#8217;s humanoid robot</a>
                with a top speed of 7 kilometers per
                hour; Honda&#8217;s ASIMO is now
                in 3rd place with a top speed of 6
                kilometers per hour.</p></description>
```

Figure 1 : example of RSS feed

Proc. of the Intl. Conf. on Advances in Image Processing and Computation Techniques - PCT 2012

ISBN: 978-981-07-2969-1

II. RELATED WORK

A personalized web search engine presented uses fuzzy concept network with link-based search method [5]. The fuzzy concept network depending on user profile which then re-order five documents' ranking with respect to user's interest. It registered five results of the link-based search engine, and the system provided personalized high quality result. Self-organizing search engine called soSpace for RSS syndicated web contents was proposed in [4]. SoSpace constructed on peer-to-peer network technology. SoSpace is capable of indexing and searching web pages described by RSS feed frequently. The experiment results showed that soSpace has good load scalability as the contents increase. The advanced and personalized search engine subsystem, which is called PeRSSonal was presented and evaluated in [6]. The researchers built PeRSSonal web portal system for the retrieval, processing, and presentation of articles. Also the system collected RSS feeds from major news portals of the Internet. Coalescence of XML-Based RSS aggregator for blogosphere was proposed in [7]. They presented a synthetic analyzer called PheRSS. The proposed analyzer aggregated different formats of RSS like atom, RSS 1.0 and RSS 2.0. A novel stream search engine, which called FeedMil is also presented [2]. For the purpose of subscription FeedMil can retrieve quality streams of topical relevance. Beyond a simple query matching, FeedMil also gives a new search experience that focused on quality and topic relevance. A personalized search engine model based on RSS's user interest was proposed [8]. The user model adopts three layers, the top layer is the user model node which represents identification number. The second layer is the channel layer and the final layer is the user interest. A self-adaptation approach to Fuzzy-Go search engine was proposed in [9]. The fuzzy search engine in each search register the difference between user's real behavior on selecting web pages and the ordering of search results. The proposed solution collected and analyzed feedbacks to adjust the fuzzy similarities between the web pages domain classification, the fuzzy ontology terms and the fuzzy factors of degrees of importance. Fuzzy Search Hash Map was presented in [10]. The presented algorithm is extension to the regular Java Hash Map data structure which allowing highly efficient fuzzy string key search. This extended hash map based on object oriented principle. It used a custom key for approximate string matching. A system of intelligent information syndication was designed [11]. The proposed system includes two parts: the first part aggregating different RSS channels information to local desktop; the second part classify the information into different categories. A fuzzy search engine based on fuzzy ontology and semantic search was also developed [12]. The developed system called fuzzy_Go. The researcher used fuzzy ontology to find similarity between terms to construct semantic search of keywords.

III. THE PROPOSED SEARCH ENGINE

A search engine model consist three main phases which are crawling using RSS feeds algorithm, indexing weblogs algorithm and searching technique with Fuzzy logic algorithm. The overview of system structure is as shown in Figure 2.

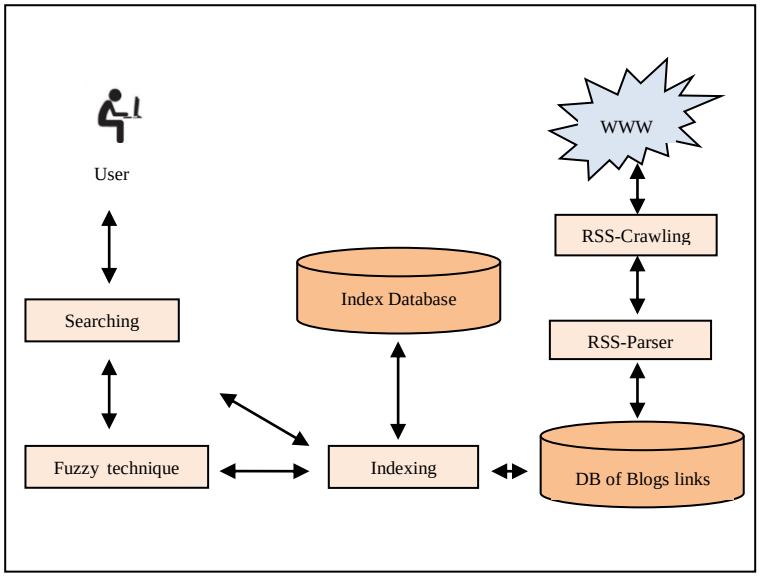

Figure 2: System structure overview

A. *RSS Feeds Crawling and Parser Algorithm*

The process begin when take link of RSS feed. Connect with Internet to retrieve the source code of RSS Feed as format XML file. Analyze and parser the file of RSS to get information of many web blog such as title, link, publish dates and description. The method of parser process done in the following steps:

- Input the xml file which is already collect from RSS crawling step.
- Read the xml nodes from.xml file by node reader.
- Check if node name is title, save title on blog database.
- Check if node name is link, save link on blog database.
- Check if node name is Date of publish , save Date of publish on blog database.
- Check if node name is description, save description on blog database.

B. *Indexing weblogs algorithm*

Proc. of the Intl. Conf. on Advances in Image Processing and Computation Techniques - PCT 2012

ISBN: 978-981-07-2969-1

Indexing weblogs involved nine steps to construct indexing database. The steps include retrieve the source code of weblog, remove HTML tags, remove digits, split the lines of string to words, remove the words that have length less than two, remove stop words, stemming the words, count the similar words and finally, build the database.

C. Searching technique with Fuzzy Parameters algorithm

Searcher is User Interface where the user type his/her queries (keywords). In this interface we use fuzzy logic technique to compute the page rank using FWA for queries. The user input two queries with different degree of importance values according to their own perceptions and feelings of the keywords that they put as queries. The importance of query is determined using linguistic variable called importance degree of query with six fuzzy sets (unimportant, rather unimportant, moderately important, rather important, very important, more important) to allow user to choose from [12]. These fuzzy sets of degree of importance of each query are mapped into six triangular fuzzy membership functions as illustrated in figure 3.

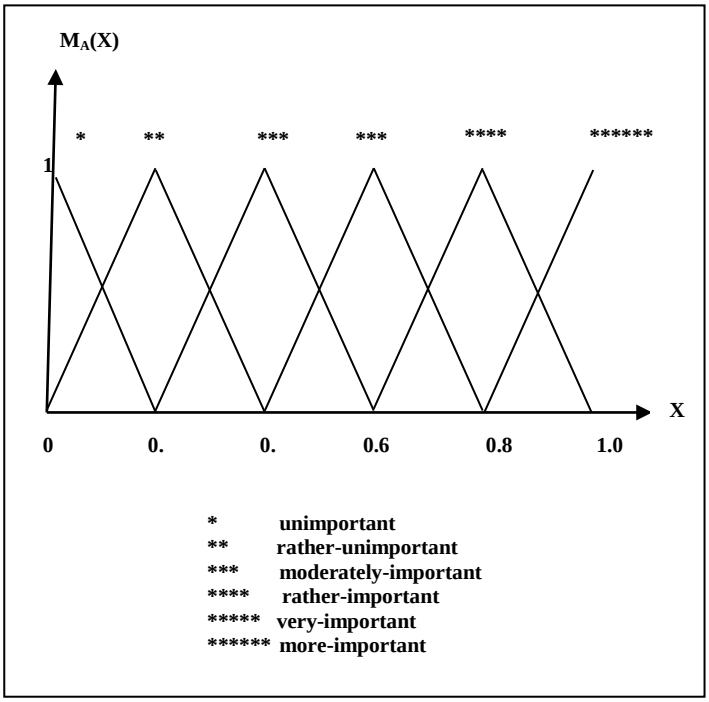

Figure 3: Membership functions that model the importance of query.

The FWA is calculated using the equation below [12]:

$$\boldsymbol{FWA(X1, X2, W1, W2) = \frac{X1W1 + X2W2}{W1 + W2}}$$

Where:

X1, X2: represent the density of query in weblog.

W1,W2: represent the importance of query.

The weblog with the highest FWA value is retrieve and treated as the first link in the query result and they are presented in the order where the highest is the first link and the lowest FWA will be the last link retrieved. Thus, it allows users to easily choose from the most relevant to the least.

IV. Experimental

The problem with searching is that there are many different measures for evaluating the performances of information retrieval systems. Every measure requires a collection of documents and queries. All common measures assume a ground truth notion of relevancy: every document is known relevant or non-relevant to specific query [13]. The most commonly used evaluation criteria are precision and recall. Precision measures the relevant documents which can only be judged by human users. For large databases, it is difficult to determine the total number of relevant blogs to calculate recall that will be too labor-intensive [14]. To solve this problem, we decided to use precision and mean average precision (MAP) instead.

Precision is defined as the fraction of the documents retrieved that are relevant to the user's query. It is defined as follows [13]:

$$\boldsymbol{Precision} = \frac{|\{\boldsymbol{retrieved\ documents}\}| \cap |\{\boldsymbol{relevent\ documents}\}|}{|\{\boldsymbol{retrieved\ documents}\}|}$$

In binary classification, precision is analogous to positive predictive value. It means how many percentages of retrieved documents are relevant to the query. Precision measures how well it is doing at rejecting non-relevant documents.

The mean average precision is another well-known evaluation criterion. It is the mean of the average precision scores for each query.

$$\boldsymbol{MAP} = \frac{\sum_{n=1}^{N} \boldsymbol{Precision(n)}}{\boldsymbol{N}}$$

Where N is the total query number.

ISBN: 978-981-07-2969-1

Ten users executed the system and tested by five queries. Each user marked the relevance of the retrieved documents of all results. The precision of each queries is calculated as is show in Figure 4 – 8.

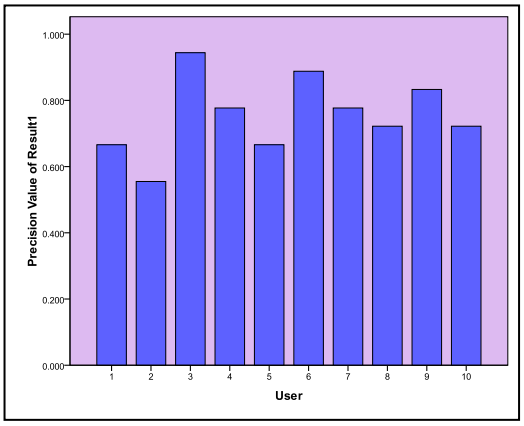

Figure 4: The precision of first result.

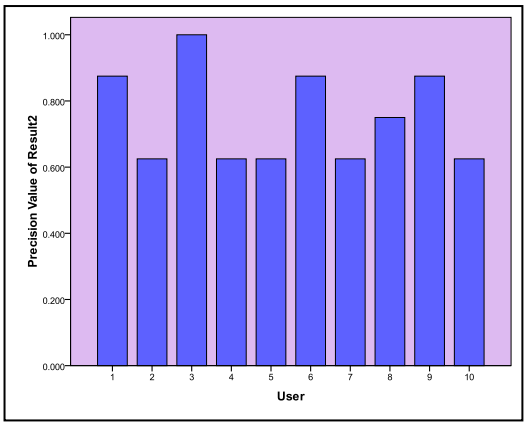

Figure 5: The precision of second result.

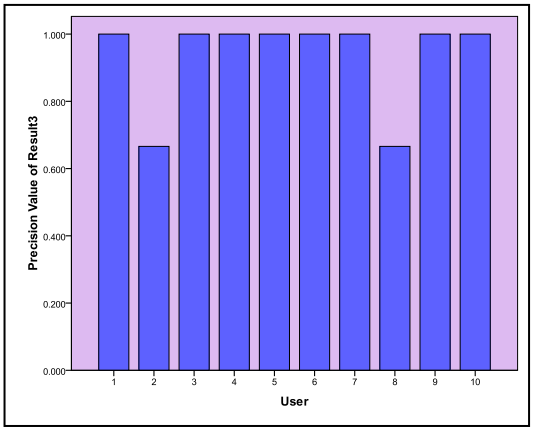

Figure 6: The precision of third result.

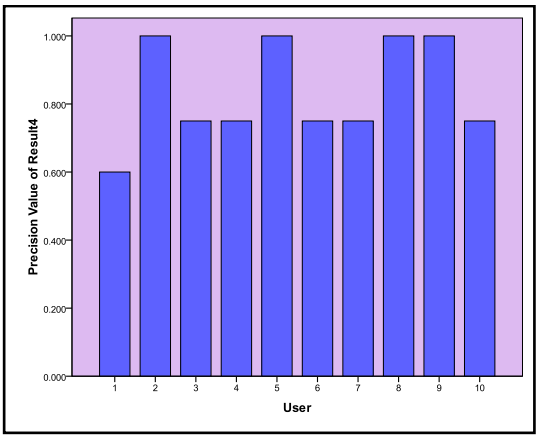

Figure 7: The precision of fourth result.

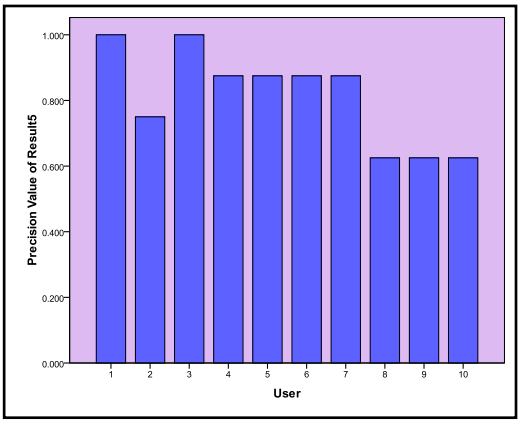

Figure 8: The precision of fifth result.

After that, the mean average precision is calculated to obtain the mean of the average precision scores for each query, which is equal to 81.7% (see Figure 9).

ISBN: 978-981-07-2969-1

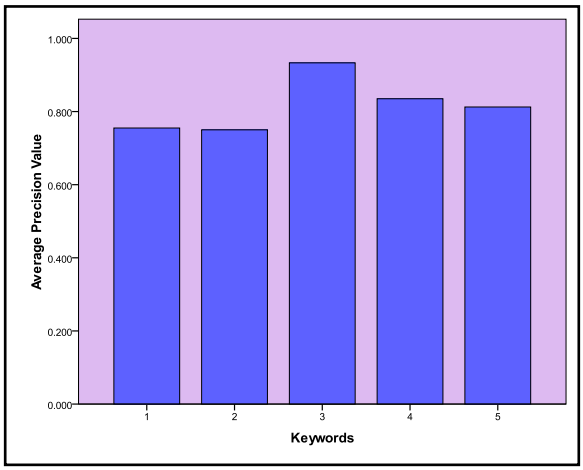

Figure 9: The mean average precision

CONCLUSION

Many people are using blogs to share their opinions and information on certain subject matters attractive to them. Sometimes, these blogs contain various valuable and rich information beside the article itself. Since there is no general rule to describe blog pages, it is more difficult to extract their contents. Thus, this study proposes blogs search engine using RSS syndication and fuzzy logic, which include three phases as mentioned. The crawling phase crawls RSS feeds and parse them over to the next phase. The crawling phase work on multi process operation that means that it crawls three RSS feeds at the same time. The indexing phase include many steps. The steps are retrieving the source code of the weblogs or blogs; removing HTML tags of the source codes; removing any digits in the source codes; splitting the lines of string into words; removing the words which length are less than two letters (for example I and a); removing stop words; stemming the words; counting the similar words; and finally building the database. The output of this phase is a database which is a two dimension array. The search phase contains three functions - calculating the frequency of keywords in document (weblog); calculating the importance of query from triangular fuzzy membership functions; and lastly calculating the fuzzy weighted average. The experiment shows a promising 81.7% mean average precision based on total system performance.

References

[1] W. Gao, Y. Tian, ,T. Huang and Q. Yang, Vlogging: a survey of videoblogging technology on the web, ACM comput. surv. 42(4), article 15 (June 2010), 57 pages

[2] J. Park, , Y. Shin, K. Kim and B. Chung, Searching the long tail of social media streams on the web, IEEE intelligent systems, 09 November 2010.

[3] G. Meghabghab and A. Kandel, Search engines ,link analysis ,and user's web behaviour, Berlin Heidelberg: Springer-Verlag, 2008.

[4] Y. Zhou, X. Chen, and C. Wang, "A self-organizing search engine for RSS syndicated web contents," Proceedings of the 22nd international conference on data engineering workshops (ICDEW'06), 24 April 2006, Atlanta, GA, USA, IEEE Computer Society.

[5] K. Kim and S. Cho, "A personalized web search engine using fuzzy concept network with link structure," IFSA world congress and 20th NAFIPS international conference, vol.1, pp. 81-86, 2001. Joint 9th.

[6] C. Bouras, V. Poulopoulos and P. Silintziris, "Personalized news search in www: adapting on user's behaviour," fourth international conference on internet and web applications and services, pp. 125-130, May 2009.

[7] T. Phoey Lee, A. Abdul Ghani, H. Ibrahim and R. Atan, "Coalescence of XML-based Really Simple Syndication(RSS) aggregator for blogosphere," Proceedings of MoMM2009, Kuala Lumpur, Malaysia, ACM, pp. 530-534, December 2009.

[8] Z. Jiang and X. Deng, "A personalized search engine model based on RSS user's interest," 2nd international conference on future computer and communication, vol. 2, pp. 196-199, 2010.

[9] Y. Lin, L. Lai, C. Wu and L.Huang, "A self-adaptation approach to fuzzy-go search engine," IEEE, pp. 1020-1025, 2010.

[10] V. Topac, "Efficient fuzzy search enabled hash map," 4th international workshop on soft computing applications, Arad, Romania, IEEE, pp. 39-44, July 2010.

[11] W. Shang, T. Wang and R. Lv, "The key technology research of intelligent information syndication," fourth international joint conference on computational sciences and optimization, pp. 865-867, 2011.

[12] L. Lai, C. Wu, P. Lin and L. Huang, "Developing a fuzzy search engine based on fuzzy ontology and semantic search," IEEE international conference on fuzzy systems, Taipei, Taiwan, pp. 2684-2689, June 2011.

[13] G. Xu, Y. Zhang and L. Li, Web mining and social networking, Techniques and application, New York, Springer, 2011.

[14] X. Zhang, C. Xu, J. Cheng, H. Lu, and S. Ma, "Effective annotation and search for video blogs with integration of context and content analysis," IEEE transactions on multimedia, vol.11(2), pp. 272-285, 2009.