Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 64 (2013) 1 – 7

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference On DESIGN AND MANUFACTURING, IConDM 2013

Comment [S1]: Els and page numbers.

Ranking of Searched Documents using Semantic Technology

Juhi Agrawal*[a],Nishkarsh Sharma[b], Pratik Kumar[c], Vishesh Parshav[d], R H Goudar[e]

Graphic Era University, Dehradun, India

juhiiagrawal@gmail.com[*],nishkarsh4@gmail.com[b],pratikkumar938@gmail.com[c],vishparshav1@gmail.com[d],rhgoudar@gmail.com[e]*

Abstract

World Wide Web is a vast resource of data growing continuously. Nowadays, it becomes increasingly hard for users to retrieve useful data due to the continually rapid growth in data volume. This vast amount of data is making search more and more difficult with traditional search engine as they return huge data for a given query which is consisting of relevant as well as irrelevant data. This is not only results in wastage of user time but also lead to data overload problem. So, users are not satisfied with searching the information by traditional search engine. So the problem of re-ranking search pages or results has become one of the main problems in IR field. Currently searching methods are mainly based on keyword matching technique but this technique has some weaknesses. The first weakness is that web users cannot express their search intention accurately or properly using several keywords. So most of the time, the exactly matched results do not satisfy the web users. Second weakness is that keyword matching cannot sure the selected candidates have high correlation with the user's query, given the different meanings of the keywords. Another problem about traditional search engines is their ranking methods. To fulfil the requirement of users we are using Semantic search engine with page ranking algorithm which will search the data semantically and holds the capability to re-rank search results effectively and try the best to arrange the web results which are most relevant for the users. The proposed algorithm for page ranking is based on result of semantic web with user attention time.

Keywords: Semantic search, semantic relevance, user profile, Term frequency.

* Juhi Agarwal. Tel.: +91-9411 39 2004
E-mail address: juhiiagrawal@gmail.com

doi:10.1016/j.proeng.2013.09.070

1. Introduction

World Wide Web is a vast resource of data growing continuously. Nowadays, it becomes increasingly hard for users to retrieve useful data due to the continually rapid growth in data volume. This vast amount of data is making search more and more difficult with traditional search engine as they return huge data for a given query which is consisting of relevant as well as irrelevant data. [1][2] This not only results in wastage of user time but also leads to data overload problem. So, users are not satisfied with searching the information by traditional search engine. So the problem of re-ranking search pages or results has become one of the main problems in IR field. Currently searching methods are mainly based on keyword matching technique but this technique has some weaknesses. The first weakness is that web users cannot express their search intention accurately or properly using several keywords. So most of the time the exactly matched results do not satisfy the web users.[3] Second weakness is that keyword matching cannot ensure the selected candidates have high correlation with the user's query, given the different meanings of the keywords[4]. Another problem about traditional search engines is their ranking methods. The main goal of our research paper is to fulfil the requirement of users with page ranking algorithm which will rank the documents in a better way and holds the capability to re-rank search results effectively and try the best to arrange the web results which are most relevant for the users using the user time attention algorithm. The proposed algorithm for page ranking is based on result of semantic web in which it will rank the pages based on term frequency factor of keywords and semantic words. Frequency factor means how many times the same keyword is repeating in the web page.

Our contribution in this paper is given below:

- Proposed Architecture for Document Ranking based on Semantic Web.
- Proposed Ranking Algorithm of Filtering Search Results Based on User Attention Time.
- Proposed Mathematical Model for ranking of searched results.

2 Related Work

Nowadays, there are a lot of re-ranking algorithms to rank the pages. Previous work has been done on the granularity of Web pages. In [5] there are three kinds of information used to re-rank documents, i.e. Document Information, Query Information and Ancillary Information but this paper is limited to appropriate algorithm. In paper [6] a new method is proposed for document re-ranking that is using inter-document relationships and expressed by distances and that can be obtained from the text. The similar work done on re-ranks in paper [7] by making clusters of the documents. In paper [8] a system is proposed that allows automatic creation of structured user profiles, which are based on an existing hierarchy. In paper [9] a system is proposed to build a user profile as a weighted concept hierarchy, which is created from the Open Directory project. In paper [10] a system proposed for re-ranking method for reordering the images retrieved from an image searching engine but it was also limited to good algorithm. In paper [11] a new approach has been proposed which assign the weights to hyperlinks, Based on its position at the page, each link gets a weight. In paper [12] [13] the web page links are weighted based on the number of in-links and out-links of their reference pages. The proposed algorithm is known as 'weighted page rank'. It was a good approach but it was limited with the accuracy because this algorithm does not ask about extra information from the user for giving an accurate ranking. In paper [14], a parameter viz. query sensitiveness technique is proposed. It measures the relevance of documents with respect to a term or topic. The web pages are ranked according to two parameters- global importance and query sensitiveness. In paper [15] the web pages are ranked on the basis of syntactic classification. This approach dose not cares about the semantics of data in the web pages. In paper [16][17] [18] a new approach is proposed. It is MFCRank ranking algorithm which is used for topic-specific search systems. The technique correlates data and creates unified link bu it was not giving good accuracy. In paper [19] [20] [21], different parameter is used with respect to a particular topic to rank the web

search result. The proposed method calculates the probability of accessing a particular page for any particular topic. So the traditional methods and papers neither were providing accuracy nor effective methods and algorithm to dynamically ranking the data according to keyword and semantics of words. Therefore, in this paper, we have adopted the architecture and algorithm for filtering the web pages or documents according to the user attention time.

3 Proposed Architecture for Document Ranking based on Semantic Web

Semantic approach is very helpful in reducing the difficulty in discerning and extracting content semantics. As shown in Fig 1, the two major tasks in the proposed semantic information retrieval system are semantic information retrieval and ranking of search results. The organization of the proposed architecture is explained below:

- User Interface: It provides the interaction between the system and the user.
- Domain Ontology: Domain ontology will contain all the information related to that particular domain.

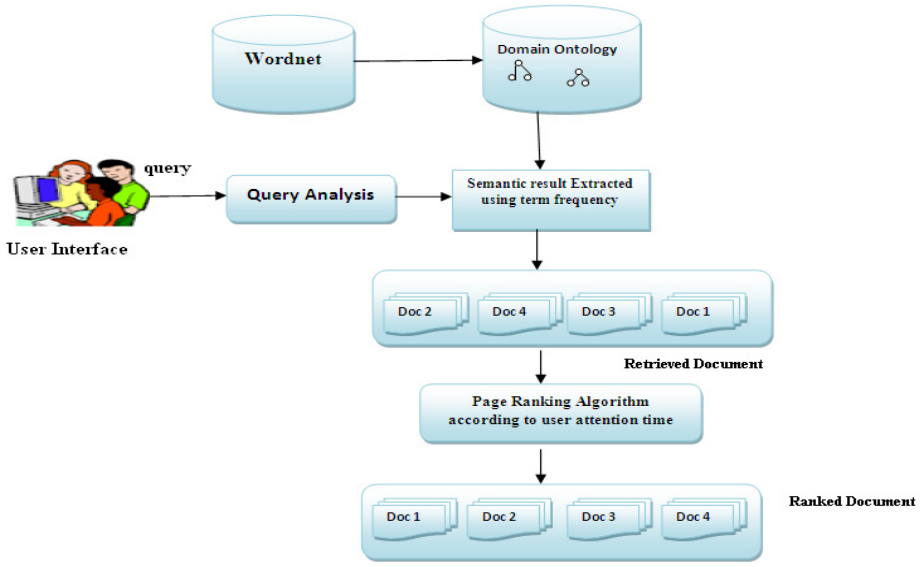

Figure 1: Architecture of the proposed Document ranking system.

- Word net: Wordnet is used for getting the synonyms of words.
- Query Analysis: The query entered by the user at user interface will be processed (stemming, change to lower case etc.) to get better results.

- Semantic result extracted using term frequency: Results would be extracted on the basis on term frequency algorithm and using synonyms of the user entered terms.

- Page Ranked Algorithm according to user attention time: User attention time algorithm would filter the results once the user starts reading a document. After each filtering, better results would be provided.

4 Proposed Ranking Algorithm of Filtering Search Results Based on User Attention Time

This algorithm depends on the user attention time on a particular document. The proposed algorithm takes care of user's reading duration for every document and if the time duration of reading that document passed a threshold value, then the system assumes that document is according to user's interest. The system will filter results and give a high rank to those documents that are similar to the documents that was having more attention earlier. These documents are sorted from high value to low value.

4.1. *User Profile:* When a user registers on the website, he is given a passage for reading to observe his reading speed to be used in User Attention Time algorithm for better results. The snapshot is given in Figure 2 for observing the reading speed of user. The user clicks on 'Start' button when he starts reading the passage and the timer start. The timer stops when the user clicks on 'Finish' button which he does when he completes reading it. The length of the document is calculated in number of words. The 'words per minute' of user is calculated by dividing the length of the document by the time taken by the user to completely read that passage. This information is stored in user profile. In case user reading speed changes then he can update his reading speed using reading passage.

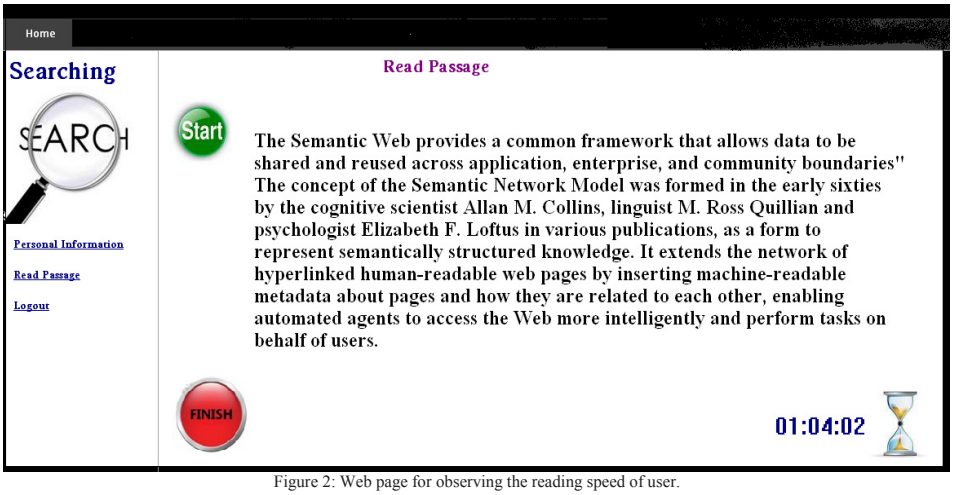

Figure 2: Web page for observing the reading speed of user.

4.2. *Search() :* The user enters the query. The query is then processed to get the individual terms or tokens. These tokens are stored in terms[]. Links to all those documents that contain any of the terms which are in the user query are returned and stored in docs[]. The total frequency of the terms that each document contains is then calculated

and stored in totalFreq[]. The list of the documents is then sorted according to decreasing term frequency and this list is passed to a function filterResult() which uses User Attention Time algorithm for filtering the results.

TABLE 1: Algorithm: Sorting of documents according to decreasing order of term frequency

```
1.  search ()
2.  {
3.      q← getUserQuery();
4.      terms[] ← processQuery(q);
5.      Docs[]← returnDocuments();
6.      for i ← 0 to n          //n  is total number of docs
7.          For j←0 to m        //m  is total number of terms found in the query
8.          {
9.                  tf [i][j]←freq(docs[i],terms[j]);
10.                 totalFreq[i]← totalFreq[i] + tf [i][j]
11.         }
12.     //Sort documents according to decreasing order of term frequency
13.     query(docs);
14. }
```

TABLE 2: Algorithm: Filtered results according to User Attention Time algorithm.

Input: Sorted list of all Documents according to term frequency returned by search ().
Output: Filtered results according to User Attention Time algorithm.

```
1.  filterResult(docs[])
2.  {
3.      listDocuments(docs[]);
                //list the documents in decreasing order of term freq
                //show the document selected by the user
4.      docLen=getDocLength(topic);
5.      wpm ← getWordsPerMinute(); //From user profile
6.      t ← docLen/wpm; // Calculate the threshold time
7.      if( hasReachedUserThreshold(t)==false)
8.      then
9.              userInterest ← false;   //The user is not interested in that document
10.     if(hasUserThresholdTime(t)==true)
11.     then
12.             userInterest ← true
13.     //bring similar documents up in the list
14. }
```

4.3. *filterResult():* This function filters the results by using the User Attention Time algorithm. The list of the documents which is passed to this function is displayed to the user. When the user clicks on a document, he is presented with that document. The length of the document is calculated in words returned by getDocLength() and stored in docLen. The 'words per minute' of the user is returned by getWordsPerMinute() function that gets the information from user's profile and stored in wpm. The threshold time is then calculated dividing docLen by wpm and stored in t. When the user starts reading the document, the timer starts and it is checked after regular interval of time if the threshold has reached. If the threshold is reached, the userInterest is set to true and the other documents that are similar to the document read by the user are moved up in the docs[] list.

5. Mathematical Model

1. Let **q** be the query entered by user
2. Let **D** denotes the set of all documents
3. Let $\mathbf{t_{fij}}$ be the term frequency of word **j** in document **i**
4. Let $\mathbf{t_{fi}}$ be the total term frequency of page **i**
5. Let $\mathbf{PR_i}$ be the rank of document **i**

$$\mathbf{t_{fi}} = \sum_{j=1}^{n} \mathbf{t_{fij}}$$ where **n** = total number of words in query **q**
$\mathbf{t_{fij}}$ = frequency of word **j** in $\mathbf{D_i}$

7. $\forall\, \mathbf{d_i}$ & $\mathbf{d_j} \in \mathbf{D}$,
8. if $\mathbf{t_f}$ of $\mathbf{d_i} < \mathbf{t_f}$ of $\mathbf{d_j} \Rightarrow \mathbf{PR_i} < \mathbf{PR_j}$
9. Let $\mathbf{L_i}$ be the length of document **i**
10. Let **S** be the reading speed of user in words per minute
11. then estimated time required to read the document may be calculated as

$\mathbf{t = L_i/S}$

$\mathbf{t_{ci}}$ is calculated as the time that has been spent by the user on document **i** and updated simultaneously while user focuses on the document, then

12. if $\mathbf{t_{ci}} = \mathbf{t}$,
13. increase **PR** of $\mathbf{D_i}$ by a single unit

6. Conclusion

We have proposed a novel concept for Document Ranking algorithm with semantic concept based on user attention time in this paper, which take use of semantic relevance and term frequency for increasing the accuracy. The designing and implementation of the algorithm are based on a set of intelligent algorithms, including semantic approach of words. This algorithm will give the better accuracy than the traditional methods.

References

[1] Guan-yu LI, Sui-ming YU and Sha-sha DAI, "Ontology based query system design and implementation", International conference on network and parallel computing, pp.1010 -1015, 2007.

[2] Jerome Euzenat, Pavel Shvaiko, "Ontology Matching", Springer-Verlag, Berlin Heidelberg(DE),2007,isbn:3-540- 49611-4

[3] Zemirli, W.N. and Tamine-Lechani, L. and Boughanem, M. 2007. "A personalized retrieval model based on inuence diagrams". Sixth International and Interdisciplinary Conference on Modeling and Using Context , 20-24 August 2007

[4] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity inWordNet," in Proceedings of ECAI, 2004.

[5] Dequan Zheng. Research on Cross-Language Information Retrieval Based on a Combination of Ontology with Statistical Language Model. Dissertation for the Doctoral Degree in Engineering, Harbin Institute of Technology, 2006: 1-3

[6] J. Balinski and C. Danilowicz. "Re-ranking method based on inter-document distances". Information Processing and Management, 41(2005), pages 759–775, 2005.

[7] V. Jain and M. Varma. "Learning to re-rank: query-dependent image re-ranking using click data". In Proceedings of the 20th international conference on World Wide Web, 2011.

[8] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li. "Video search re-ranking via multi-graph propagation". In Proceedings of the 15th international conference on Multimedia, 2007.

[9] Alani, H., and Brewster, C "Metrics for Ranking Ontologies". Proceedings of the 4th Int. Workshop on Evaluation of Ontologies for the Web (EON'06), at the 15th Int. World Wide Web Conference (WWW'06). Edinburgh, UK, 2006.

[10] W.-H. Lin, R. Jin, and A. Hauptmann. "Web image retrieval re-ranking with relevance model". In Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence, 2003.

[11] Baeza-Yates,Ricardo; Davis, Emilio; "Web page ranking using link attributes," Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, May 2004.

[12] Lee, Dick. L.; Chuang, Huei; Seamons, Kent; "Document Ranking and the Vector-Space Model;" IEEE Software, March/April 1997; pp. 67-75.

[13] Xing, W.; Ghorbani, A.; "Weighted PageRank algorithm;" Proceedings of the Second Annual Conference on Communication Networks and Services Research, 19-21 May 2004; pp. 305 – 314.

[14] Wen-Xue Tao; Wan-Li Zuo;" Query-sensitive self-adaptable web page ranking algorithm" Machine Learning and Cybernetics, 2003 International Conference on Volume 1, 2-5 Nov. 2003 Page(s):413 - 418 Vol.1

[15] Debajyoti Mukhopadhyay, Pradipta Biswas, Young – ChonKim , "A Syntactic Classification based Web Page Ranking Algorithm" , 6th International Workshop MSPT 2006.

[16] Yunming Ye1, Yan Li1, Xiaofei Xu1, Joshua Huang2, and Xiaojun Chen1,MFCRank: A Web Ranking Algorithm Based on Correlation of Multiple Features,ı c Springer-Verlag Berlin Heidelberg 2006,LNCS 3878, pp. 378–388.

[17] Zhang, L., F.Y., M., Ye, Y.: Cala: A web analysis algorithm combined with content correlation analysis method. Journal of Computer Science and Technology 18, (2003) 21–25.

[18] Gy¨ongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: VLDB 2004. Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31 - September 3, pp. 576–587 (2004).

[19] Mukhopadhyay, Debajyoti; Giri, Debasis; Singh, Sanasam Ranbir; "An Approach to Confidence Based Page Ranking for User Oriented Web Search;" SIGMOD Record, Vol.32, No.2, June 2003; pp. 28-33.

[20] Mukhopadhyay, Debajyoti; Singh, Sanasam Ranbir; "An Algorithm for Automatic Web-Page Clustering using Link Structures;" IEEE INDICON 2004 Proceedings; IIT Kharagpur, India; 20-22 December 2004; pp. 472-477.

[21] Chakrabarti, S. et. al.,; "Mining the link structure of the World Wide Web;" IEEE Computer, 32(8), August 1999.