Enhanced Retrieval of Web Pages using Improved Page Rank Algorithm

Rekha Jain [1], Sulochana Nathawat [2], Dr. G.N. Purohit [3]

[1]Department of Computer Science, Banasthali University, Jaipur, Rajasthan

[1]rekha_leo2003@yahoo.com
[2]nathawat.sulochana@gmail.com
[3]gn_purohitjaipur@yahoo.co.in

ABSTRACT

Information Retrieval (IR) is a very important and vast area. While searching for context web returns all the results related to the query. Identifying the relevant result is most tedious task for a user. Word Sense Disambiguation (WSD) is the process of identifying the senses of word in textual context, when word has multiple meanings. We have used the approaches of WSD. This paper presents a Proposed Dynamic Page Rank algorithm that is improved version of Page Rank Algorithm. The Proposed Dynamic Page Rank algorithm gives much better results than existing Google's Page Rank algorithm. To prove this we have calculated Reciprocal Rank for both the algorithms and presented comparative results.

KEYWORDS

Sense Ambiguity, Word Sense Disambiguation, Page Rank Algorithm, Evaluation Measure.

1. INTRODUCTION

Today Web is increasing very rapidly so it becomes very difficult to manage information on the Web. Therefore it is necessary for users to use efficient information retrieval techniques to get the desired information. Web Mining is the extraction and mining of useful information from the World Wide Web (WWW) [1]. Number of in-links and out-links of a web page have importance in Web Mining. Search Engines returns results that have both relevant and irrelevant information regarding the user's query. Several ranking algorithms are proposed in literature to rank web pages. Web search ranking algorithm plays an important role so that user can get the most relevant results to the user's query. For polysemous words that have several meanings, it is difficult to find relevant results at top. For the solution of this we use Word Sense Disambiguation. Word sense disambiguation is the problem of selecting a sense for a word from a set of predefined possibilities of senses. One of the important objective of WSD is that it improves the performance of various applications [2].

The structure of this paper is as follows: section 2 discusses the brief overview of Information Retrieval, section 3 describes the introduction of Word Sense Disambiguation, section 4 describes the Evaluation Methodology, section 5 presents the Page Rank Algorithm, section 6 shows the Proposed Dynamic Page Rank Algorithm, section 7 discusses the detailed overview of Results, section 8 summarizes the Conclusion. Finally references are given.

2. INFORMATION RETRIEVAL

Information Retrieval is the art of presenting, storing, organizing and accessing the information items. IR finds the documents relevant to the information need from the large document set. For searching information user can use either search engines or browse directories organized by

categories. Information Retrieval is the basic technology behind web search engine and for web users. It plays a major role to access large corpora of unstructured data. Functions of IR can be described by following:

- Text operations are applied to the text of the documents and on the description of the user information needed and transform them into a simplified form for computation.
- The documents are indexed and the index is used to execute the search.
- Searched document that meet the user query will be ranked according to their relevance.
- User will give the feedback on the retrieved document if results are not relevant. User can refine the query and restart the search process for better result.
-

Web Mining is subset of Information Retrieval [3]. Web Mining is Data Mining technique to discover the patterns from the web. Web mining consists of Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). WCM is the process of discovering the information from the web document. WSM discovers structural link between web pages. WUM identify the browsing pattern from web data through the user profile and user's behaviour [4].

3. PAGE RANK ALGORITHM

Sergey Brin and Lawrence Page developed Page Rank algorithm at Stanford University [5]. Google search engine uses Page Rank algorithm that displays the results according the user's query. Page Rank is the numeric value which represents the importance of a page on the web by simply counting the number of pages that are linking to it [1]. These links are known as backlinks. Page Rank is calculated for each page not for whole website. Page Rank represents the probability distribution over the web pages so the sum of Page Ranks of all web pages is one.

Page Rank algorithm uses following steps [6]:

(1) Calculate the Page Rank of all pages using following formula:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \quad (6)$$

Where,

PR (A) = Page Rank of page A,
PR (Ti) = Page Rank of pages Ti which links to Page A
C (Ti) = Number of outbound links on page Ti
d = Damping factor whose value lies between 0 and 1but usually its value is 0.85.

Page Rank of page A is determined by the Page Rank of those pages which links to page A using Eq. (6).

(2) Repeat step (1) until two consecutive values are same.

In Page Rank algorithm search engine return results that are ordered according to their page rank but here problem arises for polysemous words. Polysemous words have several meaning. Sense ambiguity is the problem of identifying the sense of the word. One major disadvantage of Page Rank algorithm is that page rank are calculated and stored and according to which results are indexed. It is not calculated at query time.

4. WORD SENSE DISAMBIGUATION

Word sense is the most common accepted meaning of the word. A Word Sense Ambiguity is some uncertainty about the precise Word Sense. A particular word with a particular syntactic category is associated with more than one meaning. For the solution of sense ambiguity WSD is used. WSD is the process of identifying the senses of word in textual context, when word has multiple meanings. WSD associate a word in a text or sentence having different meaning. There are two main approaches of WSD, Deep Approaches and Shallow Approaches. Deep approaches are based on world knowledge but shallow approaches do not use the world knowledge. Neighbouring words are used to identify the sense of words. Sense repository and Sense assignment are two task of WSD. WSD methods are classified into Machine learning approaches and Dictionary based approaches. In machine learning approached machine are trained to perform the task of WSD. In these approaches classifier is learned to assign fixed number of senses. In dictionary based approaches dictionary is used to retrieve all the senses of word. The sense which meets the context word is chosen as sense [2].

5. EVALUATION MEASURE

In this section we will discuss various kinds of evaluation measures of Information Retrieval techniques.

- Precision: It is the fraction of retrieved documents that are relevant. It is calculated by dividing number of relevant documents to total number of documents retrieved [7].

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (1)$$

- Recall: It is the fraction of relevant instances that are retrieved. It is calculated by dividing number of relevant documents to total number of existing relevant documents retrieved [7].

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \quad (2)$$

- F-measure: A measure which determines the weighted harmonic mean of precision and recall, called the F-measure or balanced F-score, is defined as [8]

$$F = 2.\frac{precision.recall}{precision + recall} \quad (3)$$

Here precision and recall are equally weighted so it is known as F1-measure.The F1-measure is a specialization of a general formula, the Fα -score, defined as

$$F_{\propto} = \frac{1}{\propto\left(\frac{1}{precision}\right) + (1-\propto)\left(\frac{1}{recall}\right)}$$

$$= (1+\beta^2)\frac{precision.recall}{\beta^2(precision+recall)} \quad (4)$$

Where,

$$\propto = \frac{1}{\beta^2 + 1}$$

- Reciprocal Rank: This measure is evaluated for any process that retrieves a list of response to a query ordered by probability of correctness. Reciprocal rank is the inverse of the rank of first correct answer [9].

$$reciprocal\ rank = \frac{1}{rank} \quad (5)$$

6. PROPOSED DYNAMIC PAGE RANK ALGORITHM

The aim of the proposed Dynamic Page Rank algorithm is to create a refinement layer over the existing Page Rank algorithm to resolve of ambiguity of polysemous word. Search engine like Google presents the results according to page rank of pages in decreasing order. So sometimes irrelevant results that have high Page Rank appears before the relevant results having low Page Rank. Our proposed algorithm solves this problem. When user searches for a query, results are according to Dynamic Page Rank of retrieved pages.

Proposed Dynamic Page Rank algorithm uses following steps:

(1) User enters the query.
(2) Tokenization, stemming is done on the user query then remove the stop words of query.
(3) Enhanced query is passed to the system for some search.
(4) Calculate the Dynamic Page Rank based on the Google's Page Rank.
(5) All retrieved pages are rearranged according to Dynamic Page Rank so that relevant pages appear at top of the result set.

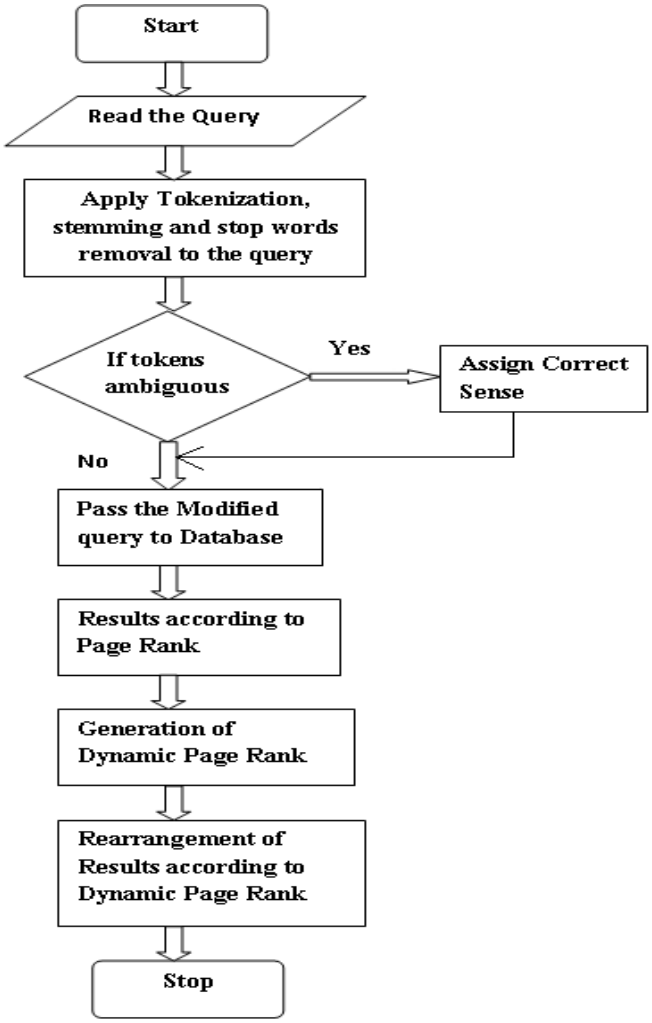

Figure 1. Flow chart for proposed Dynamic Page Rank Algorithm

7. EXPERIMENTAL RESULTS

When the user enters the word String, to be searched, Google results returns the pages for String(Program) and String(Jewellery) but the String(Program) pages are on the top priority and user need to traverse result pages for finding the String(Jewellery). The results of Google are shown in fig 2. The results produced by our algorithm were much efficient as when we applied the measures on both the algorithms. We have compared both the algorithms (Page Rank and newly developed algorithm) on the basis of Reciprocal Rank.

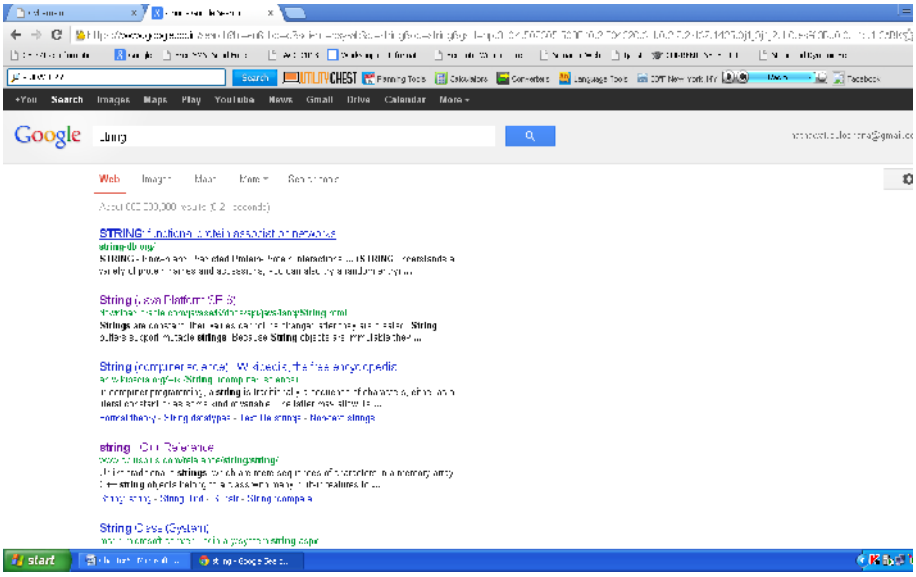

Figure 2. Google Result

Fig. 3 shows the results when user searches for string in sense of program.

Figure 3. When user searches for String (Program)

Fig. 4 shows the results when user searches for string in sense of jewellery.

Figure 4. When user searches for String (Jewellery)

Fig. 5 shows a graph for the ambiguous word string to compare Reciprocal Rank values of Page Rank Algorithm and Proposed Dynamic Page Rank Algorithm.

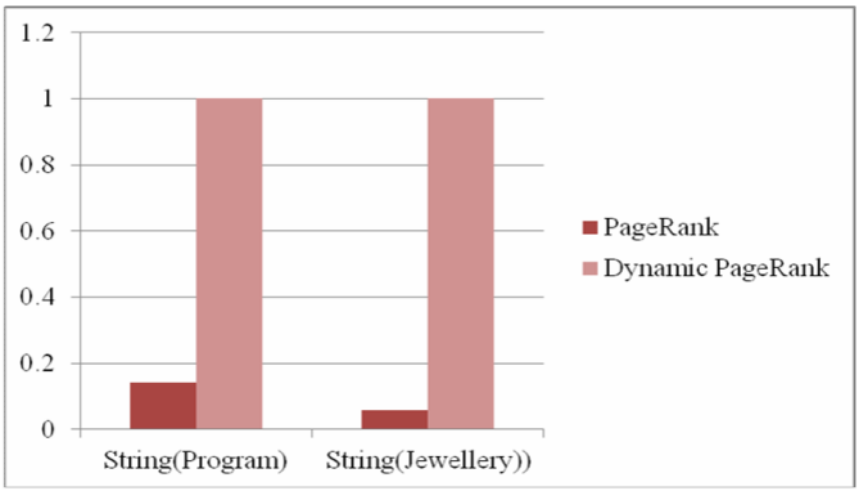

Figure 5. Comparative graph for reciprocal rank

8. CONCLUSIONS

Our proposed algorithm resolves the ambiguity of polysemous words and presents the results according to user preferences. Results shows that proposed Dynamic Page Rank algorithm is more efficient than existing Page Rank algorithm.

REFERENCES

[1] Pooja Sharma. Deepak Tyagi, "Weighted Page Content Rank for Ordering Web Search Result", International Journal of Engineering Science and Technology, Vol. 2, No. 12, pp. 7301-7310, 2010.

[2] Rekha Jain, Sulochana Nathawat, "Sense Disambiguation Techniques: A Survey", International Journal of Advances in Computer Science and Technology, Vol. 1, No. 1, pp. 1-6, 2012.

[3] Diana Inkpen, "Information Retrieval on the Internet", PhD thesis, University of Toronto.

[4] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[5] Google PageRank – Algorithm available at http://pr.efactory.de/e-pagerank-algorithm.shtml.

[6] Hema Dubey, Prof. B. N. Roy, "An Improved Page Rank Algorithm based on Optimized Normalization Technique", International Journal of Computer Science and Information Technologies, Vol. 2, No. 5, pp. 2183-2188, 2011.

[7] Precision and recall available at http://en.wikipedia.org/wiki/Precision_and_recall..

[8] R. Navigli, "Word Sense Disambiguation: a Survey", ACM Computing Surveys, Vol. 41, No. 2, pp. 1-69, 2009.

[9] Mean reciprocal rank available at http://en.wikipedia.org/wiki/Mean_reciprocal_rank.

About The Authors

Rekha Jain completed her Master Degree in Computer Science from Kurukshetra University in 2004. Now she is working as Assistant Professor in Department of "Apaji Institute of Mathematics & Applied Computer Technology" at Banasthali University, Rajasthan and pursuing Ph.D. under the supervision of Prof. G. N. Purohit. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has various National and International publications and conferences.

Sulochana Nathawat is pursuing her M.Tech degree in Computer Science and Engineering from Banasthali Vidyapith, Rajasthan. She received Master Degree in Computer Application from Apex Institute of Management & Science, Jaipur, Rajasthan in 2010. Her research interest includes Web Mining, Data Mining, Semantic Web, Information Retrieval and Natural Language Processing.

Prof. G. N. Purohit is a Professor in Department of Mathematics & Statistics at Banasthali University (Rajasthan). Before joining Banasthali University, he was Professor and Head of the Department of Mathematics, University of Rajasthan, Jaipur. He had been Chief-editor of a research journal and regular reviewer of many journals. His present interest is in O.R., Discrete Mathematics and Communication networks. He has published around 40 research papers in various journals.