ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

Efficient Information Retrieval for Ambiguous Words

Rekha Jain[1], Sulochana Nathawat[2], Rupal Bhargava[3], G.N. Purohit[4]

Department of Computer Science, Banasthali University, Jaipur, Rajasthan, India[1, 2, 3, 4]

Abstract: Information Retrieval is the task of representing, storing, organizing and offering access to information items. To retrieve required information from Word Wide Web, search engines performs number of activities. When user enters the query on to the interface of search engine he/she can get irrelevant documents if the search term contains ambiguous keywords. On web there exist numbers of ranking algorithm to retrieve relevant information. Most popular search engine Google uses Page Rank algorithm to retrieve the results. The results are arranged from higher to lower page rank values. This paper proposes an algorithm for efficient retrieval of information on the Web. It helps in word sense disambiguation and improves the performance of system. To prove the efficiency of our proposed algorithm we have applied two measures Mean Reciprocal Rank and Mean Average Precision.

Keywords: Ambiguity, Word Sense Disambiguation (WSD), Information Retrieval (IR), Web Mining, Data Mining.

I. INTRODUCTION

Internet has huge amount of information in the form of text, audio, video and other kind of document. Still there is large part of internet that is not accessible by the user. User searches for information either in search engine or browse the directories organized by categories. There are various kinds of search engines available on the Web like Google, AltaVista, Yahoo etc. user enters the keywords in the search engine that describes the information need [1]. Information Retrieval is the process of finding documents that satisfies the information need from large collection of documents. An ambiguous term is a word having multiple senses and sense can be identified by the context in which it is used. Word sense disambiguation is an attempt to remove ambiguity of polysemous words or phrases. WSD heavily relies on knowledge. Various kinds of rules are used for association of meaning in textual context.

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

II. INFORMATION RETRIEVAL

Search Engines have two main components: Web crawler and Information Retrieval (IR) system. Web crawler collects Web pages for IR system. An Information Retrieval system is a software programme that stores and manipulates information on documents. There are three basic processes of information retrieval system: the presentation of the content of the documents, the representation of the user's information need and the comparison of the two representations. Information Retrieval architecture is depicted in fig. 1 [1].

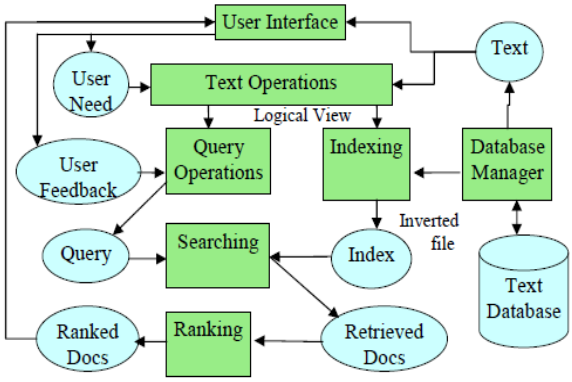

Fig. 1 Information Retrieval System Architecture

User interface manages interaction with the user. Text operations include tokenization, stemming, stop word removal. Indexing creates the inverted index of word to document pointer. Query operations transform the query for improved retrieval. Searching retrieves the documents that contain keywords from the inverted index. Ranking assign ranks to retrieved documents according to their relevancy. After retrieval of ranked document user can give feedback for refinement of query and restart the search for better results.

III. DATA MINING

Data mining is also known as Knowledge Discovery in Databases (KDD). Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information. It is the part of the knowledge discovery process. This process of extracts or mines large amount of data. It is the natural evolution of information technology. In data mining concern is about the development of methods and techniques to identify sense of data. The data mining tools appeared with the intension of facilitating data analysis and visualization as well as the discovery of useful information for decision

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

making [2]. Data Mining task is classified into two categories: descriptive and predictive. Descriptive mining describe the general properties of data in database. Predictive mining perform inference on the current data to make predictions [3].

IV. WEB MINING

Web Mining is the application of Data Mining to extract information from Web data. Web data consist of Web Content, Web Structure, and Web Usage data. fig. 2 shows the complete process of extracting information from Web data [4].

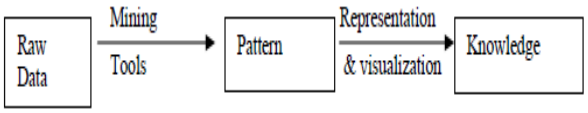

Fig. 2 Web Mining process

Web Mining process is explained as:

(1) First step is to find the resource that is to search for intended Web document.

(2) Then from retrieved resources required information is selected from retrieved resources and pre-processing is done.

(3) Generalization is done that is automatically discovers patterns at individual Web site as well multiple Web sites.

(4) Validation and interpretation of mined patterns is done.

Web mining taxonomy is depicted into fig. 3 [4]. According to Fig. 3 Web Mining can be categorized into following three types.

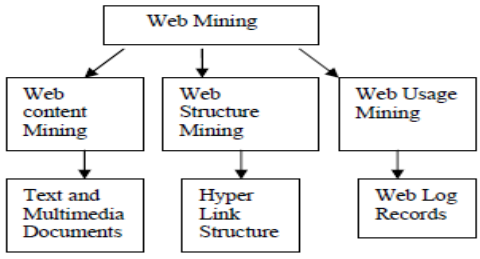

Fig. 3 Web Mining Taxonomy

A. Web Content Mining (WCM)

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data consist of text, images, audio, video etc. Web content mining is related to data mining because many data mining techniques can be applied to the web content mining. It is also related to text mining because mostly web contents are text based.

B. Web Structure Mining (WSM)

Web Structure Mining is the process of extracting structural information from Web documents. Structure of the web can be easily understood with the concept of web graph. Web graph consists of nodes and edges where Web pages act as nodes and links between pages are treated as edges. Its task is to generate structure summary about the Web site and Web page.

C. Web Usage Mining (WUM)

Web Usage Mining is the process of extracting meaningful pattern from Web documents. Usage data extract the identity, behaviour of Web users. One important challenge to web usage mining application is that Web server log data are anonymous and can create problem in identifying users and user sessions from the data.

V. WORD SENSE DISAMBIGUATION

A word, phrase or sentence is ambiguous if it has more than one meaning. Lexical ambiguity comes from the homonymy and polysemy. Homonymy of ambiguous word has same pronunciation but having different meaning. Bark of a dog verses the bark of tree is an example of homonymy. While polysemy of ambiguous word has same pronunciation having two or more distinct but related meaning. Tear for eye and tear for rip is example of polysemy. In this paper we concern about the lexical ambiguity.

Word Sense Disambiguation is the process of association of given word with meaning (sense) in context that is distinguishable from other meaning of the word. In WSD first all senses of all words are identified and appropriate sense is assigned to each occurrence of word in textual context.

WSD is a key problem of Natural Language Processing (NLP). It is used in many applications like Machine Translation (MT), Information Retrieval (IR), Speech Processing (SP), Information Extraction (IE) [5] etc.

VI. PROPOSED ALGORITHM

Google search engine uses Page Rank algorithm to retrieve the results according the user's need. In most of the cases Page Rank algorithm provides fruitful results but in case of polysemous it never considers ambiguities that lie between words and finally user gets irrelevant contents at the top of search result. Relevancy is defined as the condition of being relevant. Our proposed algorithm (Dynamic Page Rank algorithm) acts as a layer on to search engine. This algorithm resolves the ambiguities of polysemous words. In proposed algorithm firstly the enhancement of query is performed by applying various steps like Tokenization, Stemming, Stop word removal as well as sense disambiguation approach used to disambiguate the sense. Now the enhanced query string is passed to search engine to perform some search. After receiving the result in form of web pages, dynamic page rank of each page is calculated separately on the basis of matching process of matching process of enhanced search terms. This matching process matches the search terms with the information regarding web pages that is stored in web database. Finally the result is rearranged from higher to lower dynamic page rank values. It allows the user to

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

receive more meaningful contents at the top of search results. Fig. 4 shows the flowchart of proposed Dynamic Page Rank algorithm.

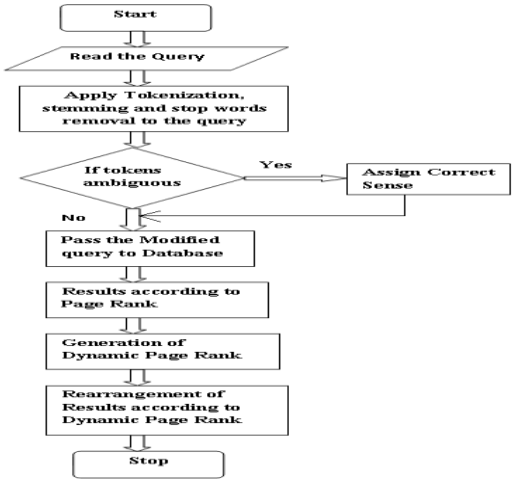

Fig.4 Flowchart of Proposed Algorithm

VII. EXPERIMENTAL RESULTS

We have applied Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) to compare the efficiency of Page Rank algorithm and Dynamic Page Rank algorithm. Mean Reciprocal Rank is a statistical measure for evaluating any process that produces a list of possible responses to a query, ordered by probability of correctness. Reciprocal rank is the inverse of the rank of first correct answer.

$$Reciprocal\ Rank = \frac{1}{rank} \qquad (1)$$

Mean Reciprocal Rank is the average of the reciprocal ranks of results for a sample of queries Q [6]:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \qquad (2)$$

Precision is the fraction of retrieved documents that are relevant and recall is the fraction of relevant instances that are retrieved. A measure that uses both precision and recall is the average precision. Precision as a function of recall is denoted by p(r). Average Precision computes the average value of p(r) over the interval r=0 to r=1.

$$AveP = \sum_{k=1}^{n} p(k)\,\Delta r(k) \qquad (3)$$

Where, k is the rank in the sequence of retrieved documents, n is the number of retrieved documents. Mean average precision for a set of queries is the mean of the average precision scores for each query.

$$MAP = \frac{\sum_{q=1}^{Q} AveP}{Q} \qquad (4)$$

Where, Q is the number of queries [7]. For example, suppose user searches for four ambiguous sample queries Tear (Eye), Tear (Rip), Sink (Basin) and sink (Technology). Dynamic Page Rank algorithm gives the first as the most correct answer. Our system calculates the reciprocal rank and average precision for these three queries using queries (1) and (3) respectively. Then calculate the Mean Reciprocal Rank and Mean Average Precision using queries (2) and (4) formulas respectively.

TABLE I

RECIPROCAL RANK OF PAGE RANK ALGORITHM AND PROPOSED ALGORITHM

Query	**Reciprocal Rank (Page Rank Algorithm)**	**Reciprocal Rank (Proposed Dynamic Page Rank Algorithm)**
Tear (Eye)	1	1
Tear (rip)	0.1429	1
Sink (Basin)	0.5	1
Sink (Technology)	1	1

TABLE II

AVERAGE PRECISION OF PAGE RANK ALGORITHM AND DYNAMIC PAGE RANK ALGORITHM

Query	**Average Precision (Page Rank Algorithm)**	**Average Precision (Dynamic Page Rank Algorithm)**
Tear (Eye)	0.5173	0.6729
Tear (rip)	0.3344	0.5527
Sink (Basin)	0.4870	0.6647
Sink (Technology)	0.4677	0.6536

On the basis of Reciprocal Rank and Average Precision of three queries we calculate the Mean Reciprocal Rank and Mean Average Precision of three queries using formula 2 and 4.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

TABLE III

MRR AND MAP OF PAGE RANK ALGORITHM AND DYNAMIC PAGE RANK ALGORITHM

	Page Rank Algorithm	**Dynamic Page Rank Algorithm**
Mean Reciprocal Rank	0.661	1
Mean Average Precision	0.452	0.636

Following fig. 5 depicts the comparative results of Mean Reciprocal Rank and Mean Average Precision of Google's Page Rank algorithm and Dynamic Page Rank algorithm.

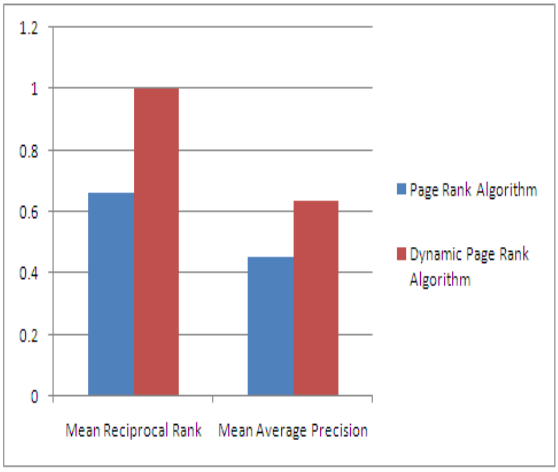

Fig. 5 Comparative Results of MRR and MAP

VIII. CONCLUSION

In Page Rank algorithm search engine return results that are ordered according to their page rank but here problem arises for polysemous words. Proposed algorithm helps in resolving the ambiguity. Results show that Dynamic Page Rank Algorithm gives more efficient results than Existing Google's Page Rank algorithm.

REFERENCES

Diana Inkpen, "Information Retrieval on the Internet

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, "From Data Mining to Knowledge Discovery in Database", AI Magazine 17 (3): 37-54 (1996).

[2] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, 2nd ed., 2006.

[3] Pooja Sharma, Deepak Tyagi, Pawan Bhadana, "Weighted Web Content Rank for Ordering Web Search Result", International Journal of Engineering Science and Technology, Vol. 2 (12), 2010, 7301-7310 .

[4] 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.

[5] Mean reciprocal rank. [Online]. Available: http://en.wikipedia.org/wiki/Mean_reciprocal_rank.

[6] Information retrieval. [Online]. Available: http://en.wikipedia.org/wiki/Information_retrieval.

Biography

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.

Rupal Bhargava is pursuing her M.Tech in Computer Science from Banasthali Vidyapith, Rajasthan. She is undergoing the training of her M.Tech in supervision of Mrs. Rekha Jain. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has published various papers in the conferences and journals.

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.