

MODIFIED PAGE RANK ALGORITHM TO SOLVE AMBIGUITY OF POLYSEMIOUS WORDS

Rekha Jain¹, Sulochana Nathawat², G.N. Purohit³

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

¹rekha_1eo2003@yahoo.com

²nathawat.sulochana@gmail.com

³gn_purohitjaipur@yahoo.co.in

ABSTRACT

Ambiguity in natural language has an effect on Information Retrieval (IR) system in general and web search system in particular. Word sense ambiguity is the reason behind the poor performance of IR system. If the ambiguous words are correctly disambiguated then IR performance can increase. WSD is defined as the task of identifying the sense of word in textual context. Word Sense Disambiguation (WSD) is the central problem of language processing. WSD improves the IR in many ways. Current IR system do not use explicit WSD and rely on user typing enough content in the query to only retrieve documents relevant to the intended senses. This paper proposes an algorithm for efficient retrieval of information on the Web according to user's need. Results are more accurate and they are according to user's preference.

KEYWORDS

Polysemous words, Ambiguity, Word Sense Disambiguation (WSD),, Information Retrieval (IR).

1. INTRODUCTION

Ambiguity of a word is to have more than one interpretation. Lexical ambiguity can be divided into homonymy and polysemy. Polysemy of a word is having multiple related meaning. Homonymy of a word is having same spelling or same sound but have different senses. Bark of a dog verses the bark of tree is an example of homonymy. Box for container and box for theatre is an example of polysemy. In this paper we are dealing with noun polysemy words. If a query contains a polysemous word then for one meaning the precision will be affected. If query contain the relevant sense of word then it improve precision. Context is a way for resolving the ambiguity. Context has the major role in identifying the meaning of polysemous word. Word Sense Disambiguation (WSD) concerns about the context of target word so that correct information for its disambiguation can be provided. Disambiguation is the process of resolving the ambiguity when a word is ambiguous. In the field of computational linguistic the problem of lexical ambiguity is known as WSD. WSD is the ability to identify the intended meaning of a word in context. It is a key problem of Natural Language Processing (NLP). People either browse or use the search service when they want to find specific information on the Web. When user searches through service he or she inputs a simple keyword in the form of query and the query response is the list of pages ranked on the basis of their similarity to the query. The basic function of any search engine is to crawl on the web based on the user query and retrieve the results to the end users. Information Retrieval (IR) is the process of obtaining the relevant information

information according to user's information need then document is relevant. A perfect retrieval system would retrieve only relevant documents and no irrelevant documents. [1].

The structure of this paper is as follows: section 2 provides the brief overview of Information Retrieval, section 3 describes the introduction of Word Sense Disambiguation, section 4 describes the Role of WSD in IR, section 5 discusses the approaches of WSD, section 6 shows the Proposed Dynamic Page Rank Algorithm, section 7 discusses the results, section 8 summarizes the Conclusion.

2. INFORMATION RETRIEVAL

An Information Retrieval system is a software programme that stores and manipulates information on documents. The system gives the information in the form of document required by the user. The documents that contain the solution for user's need are known as the relevant documents. When a user searches for a query system returns both relevant and non relevant documents. For two users the query relevancy criteria of retrieved document may differ. There are three basic processes of information retrieval system:

- 1) The presentation of the content of the documents,
- 2) The representation of the user's information need and
- 3) The comparison of the two representations.

These processes are depicted through the Figure 1.

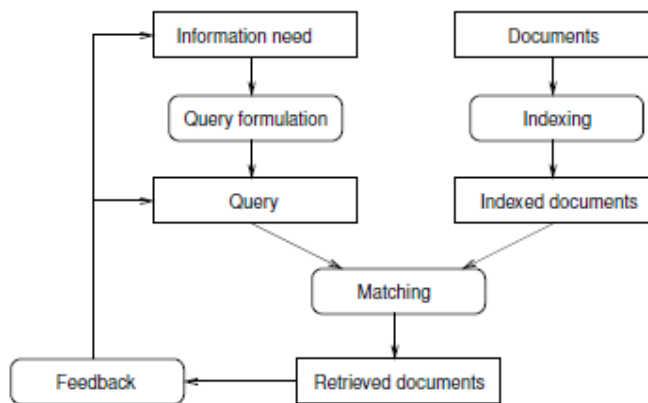


Figure 1. Information Retrieval Processes

Query formulation process is the representation of user's information need. This representation is called the query. Indexing process is the representation of the documents. In matching process query is compared against the indexed documents. This process gives the ranked documents where relevant documents are at the top of result list. User searches for required information and gives the feedback of the retrieved documents if results are not relevant. User can refine the query and restart the search process for better results [1].

3. WORD SENSE DISAMBIGUATION

A word, phrase or sentence is ambiguous if it has more than one meaning. There are three types of ambiguity: Lexical ambiguity, syntactic ambiguity and semantic ambiguity. Lexical ambiguity comes from the existence of homonymy and polysemy. Homonymy of ambiguous word has same pronunciation but having different meanings. For example, night and knight are pronounced same but have different meanings like night for period of darkness and knight for a mounted soldier. While polysemy of ambiguous word has same pronunciation having two or more distinct but related meaning. For example, down is ploysemy word having different senses like American football, soft furry feather etc. Syntactic ambiguity is when a sentence has more than one meaning based on the syntax of sentence. Semantic ambiguity occurs when a sentence have ambiguous words or phrase that have more than one meaning [2]. 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 attributable to that word. WSD task involves two steps :

1. Sense Repository: Identify all the different meanings of all the words relevant to the text under consideration. It may be from list of senses in dictionaries, from synonyms in thesaurus, from translations in a translation dictionary.
2. Sense Assignment: It involves the assignment of appropriate sense to each occurrence of word in textual context.

Context of target word gives information for its disambiguation. Context has major role in identifying the meaning of polysemous words. Context is used in two ways: Bag-of-words approach and Relational information. In Bag-of-word approach content is considered as words surrounding the target word. In this, context does not have any kind of relationship to the target word in terms of distance, grammatical relations etc. In relational information approach context is considered as some kind of relation to the target in terms of distance, syntactic relations etc.

WSD used in many applications like Machine Translation (MT), Information Retrieval (IR), Speech Processing (SP), Information Extraction (IE) [3] etc.

4. ROLE OF WSD IN IR

WSD has the detrimental effect on the precision of the text based IR. Keyword based IR systems face the problem when there is variance meaning of word. We need two assumptions for word sense: only one sense of word is activated at each occurrence and each word has a fixed number of senses. IR system will increase the performance if the documents are retrieved based on the word senses rather words. Researches were conducted to investigate this method of the representation of the documents. First time disambiguation used with IR system was done by the Weiss [4] and reported as 1% improvement in IR performance. Krovetz and Croft [5] also performed research on the relation of WSD and IR and reported that WSD did not have more impact on IR but beneficial to IR if some words are common between query and the document. Voorhees [6] applied disambiguation to IR using WordNet but resulted in drop in IR performance [7].

5. WSD APPROACHES

There are four conventional approaches of WSD [8] [9]:

5.1. Dictionary and Knowledge Based Methods

Dictionary and Knowledge based methods uses the dictionaries, thesauri, ontologies etc to retrieve different senses of word in context. Main knowledge based techniques are: the overlap of sense definition, selectional restrictions and structural approaches. Most approaches use the WordNet as sense inventory.

5.2. Supervised Methods

Supervised approaches use the sense-annotated corpora and generate the classifier system. Classifier is used for the classification of word to assign appropriate sense to each instance of that word. The training set is used to learn the classifier. Target word is manually tagged with sense from sense inventory. Supervised methods give better results than unsupervised approaches.

5.3. Semi-Supervised Methods

Semi-Supervised methods allow both labelled and unlabeled data. These use small annotated corpus as seed. Seed is used to train initial classifier using any supervised method. Initial classifier extract large training set from the remaining untagged corpus. Thus on repeating this process we will get a series of classifier until the whole corpus is consumed or given maximum number of iteration is reached.

5.4. Unsupervised Methods

Unsupervised methods work on raw unannotated corpora. Word senses are induced from input text by clustering word occurrences and then classifying new occurrences into the induced clusters. These methods do not use any dictionaries, thesauri, ontologies etc. In these sense is labelled externally to target word. The occurrence of word is divided into the number of classes by checking for any two occurrences whether they belong to same sense or not [Schutze 1998, page 97] . The task in these methods is to identify the sense clusters.

6. DYNAMIC PAGE RANK ALGORITHM

Our proposed work i.e. Dynamic Page Rank algorithm is an extension of the existing Page Rank algorithm. Page Rank algorithm presents the results sorted on the basis of Page Rank. All results of Page Rank algorithm are not relevant according to user's need. But our algorithm presents the results based on Dynamic Page Rank and highly relevant results will be retrieved to the user. This algorithm resolves the ambiguity of polysemous words. Figure 2 shows the flowchart of proposed Dynamic Page Rank algorithm.

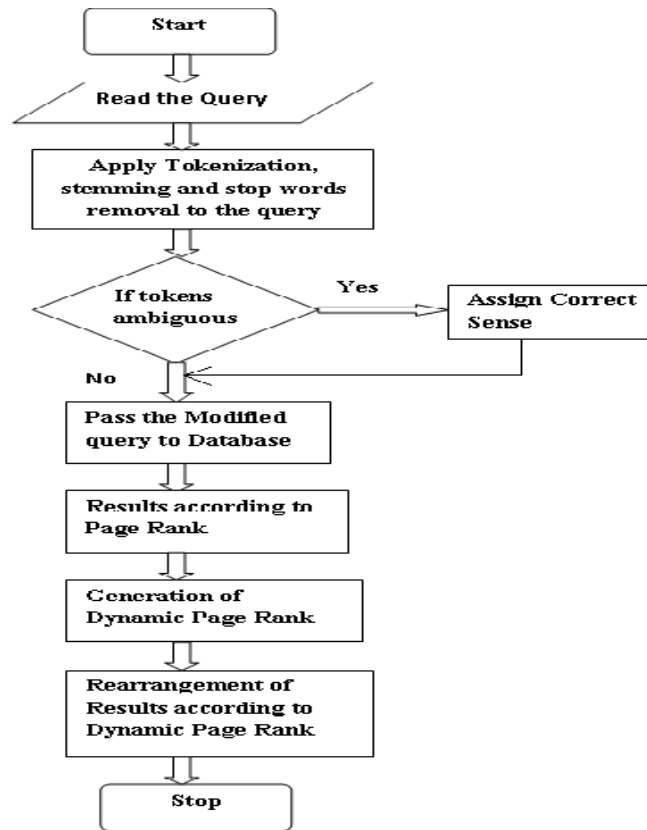


Figure 2. Flowchart of Proposed Algorithm

7. EXPERIMENTAL RESULTS

We have applied Mean Reciprocal Rank to compare the efficiency of Page Rank algorithm and Proposed 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.

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

Mean Reciprocal Rank is the average of the reciprocal ranks of results for a sample of queries Q [10]:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \quad (2)$$

For example, suppose user searches for three sample queries down (football), box (container), and command (computer). Our proposed algorithm gives the first as the most correct answer. Our system calculates the reciprocal rank for these three queries and then calculates the Mean Reciprocal Rank using above formulas.

Table 1. Reciprocal Rank of Page Rank Algorithm and Proposed Algorithm

Query	Reciprocal Rank (Page Rank Algorithm)	Reciprocal Rank (Proposed Dynamic Page Rank Algorithm)
Down (Football)	0.5	1
Box (Container)	0.25	1
Command (Computer)	0.2	1

On the basis of Reciprocal Rank of three queries we calculate the Mean Reciprocal Rank of three queries using formula 2.

Table 2. Mean Reciprocal Rank of Page Rank Algorithm and Proposed Algorithm

	Page Rank Algorithm	Proposed Dynamic Page Rank Algorithm
Mean Reciprocal Rank	0.3167	1

Following Figure 3 depicts the comparative results of Mean Reciprocal Rank of Google’s Page Rank algorithm and Proposed Dynamic Page Rank algorithm.

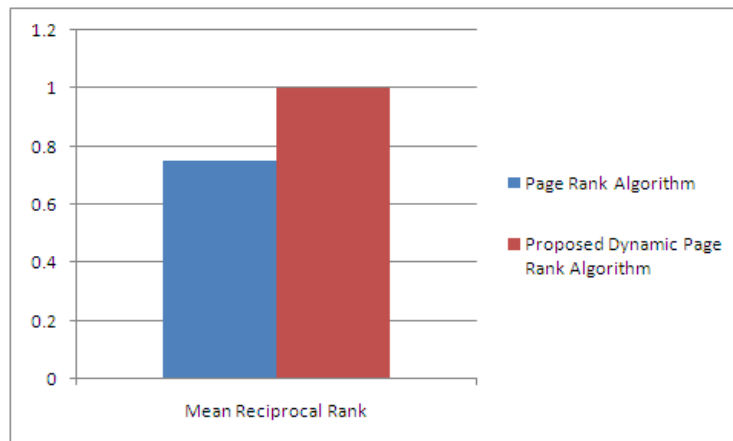


Figure 3. Comparative Results of Mean Reciprocal Rank

8. CONCLUSION

Page Rank algorithm is widely used and most accepted algorithm. But it can’t resolve the lexical ambiguity of polysemous words, phrases etc. Proposed algorithm helps in resolving the ambiguity. Results show that Dynamic Page Rank Algorithm gives much better results than Existing Google’s Page Rank algorithm.

REFERENCES

- [1] Djoerd Hiemstra, "Information Retrieval Models", In: Ayse Goker and John Davies (eds.), Information Retrieval: Searching in the 21st Century, Wiley, 2009.
- [2] Ambiguity available at <http://online.sfsu.edu/kbach/ambguity.html>.
- [3] Nancy Ide, Jean Veronis, "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art", Computational Linguistics. 24(1): 2-40, 1998.
- [4] Weiss SF, "Learning to disambiguate", Information Storage and Retrieval, 1973; 9:33-41.
- [5] Krovetz R, Croft WB, "Lexical Ambiguity and Information Retrieval", ACM Transactions on Information Systems, 1992: 10.
- [6] Voorhees EM, "Using WordNet to disambiguate word sense for text retrieval", Proceedings of ACM SIGIR Conference, 1993;16:171-180.
- [7] Liqi Gao, Yu Zhang, Ting Liu, Guiping Liu, "Word Sense Language Model for Information Retrieval", AIRS 2006: 158-171
- [8] Word-sense disambiguation available at http://en.wikipedia.org/wiki/Word-sense_disambiguation.
- [9] 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.
- [10] 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.

