AN ENHANCED APPROACH TO OPTIMIZE WEB SEARCH BASED ON PROVENANCE USING FUZZY EQUIVALENCE RELATION BY LEMMATIZATION

Divya*

Tanvi Gupta*

ABSTRACT

In this paper, the focus is on one of the pre-processing technique i.e. stemming, instead of this, 'lemmatization' can be used which is more robust than stemming as it often involves usage of vocabulary and morphological analysis, as opposed to simply removing the suffix of the word. As, World Wide Web users use search engines for retrieving information in web . But, the problem is that performance of a web search is greatly affected by flooding of search results with information that is redundant in nature i.e., existence of near-duplicates. Such near-duplicates holdup the other promising results to the users. Many of these near-duplicates are from distrusted websites and/or authors who host information on web. Such near-duplicates may be eliminated by means of Provenance, and this web search technique has been optimized by fuzzy clustering method which is based upon fuzzy equivalence relation called fuzzy equivalence clustering method.

Keywords: *near-duplicates, lemmatization , Provenance, fuzzy clustering , fuzzy equivalence relation.*

*Lingaya's university, Faridabad, India.

1. INTRODUCTION

There is a tremendous growth of information on World Wide Web over the last decade. The new challenges are created by Web for information retrieval as the amount of information on the web and number of users using web growing rapidly. It is practically impossible to search through this extremely large database for the information needed by user. Hence the need for Search Engine arises. Search Engines uses crawlers to gather information and stores it in database maintained at search engine side. For a given user's query the search engine searches in the local database and very quickly displays the results.

However, in any web search environment there exist challenges when it comes to providing the user with most relevant, useful and trustworthy results, as mentioned below:

• The lack of semantics in web

• The enormous amount of near-duplicate documents

• The lack of emphasis on the trustworthiness aspect of documents

There are also many other factors that affect the performance of a web search. Several approaches have been made and still researches are going on to optimize the web search.

A. Clustering method based on fuzzy equivalence relation:-

Web Mining has fuzzy characteristics, so fuzzy clustering is sometimes better suitable for Web Mining in comparison with conventional clustering. Fuzzy clustering is a relevant technique for information retrieval. As a document might be relevant to multiple queries, this document should be given in the corresponding response sets, otherwise, the users would not be aware of it. Fuzzy clustering seems a natural technique for document categorization. A Clustering method based upon fuzzy equivalence relations is being proposed for web document clustering. The downloaded documents and the keywords contained therein and stored in a database by the crawler. The indexer extracts all words from the entire set of documents and eliminates words i.e. stop words such as “a”, “and”, “the” etc from each documents. These keywords fetch the related documents and stored in the indexed database. The documents are stored in indexed database based on keywords. Now, the fuzzy clustering method based upon fuzzy equivalence relations [1] is applied on the indexed database.

B. ***Provenance***

One of the causes of increasing near-duplicates in web is that the ease with which one can access the data in web and the lack of semantics in near-duplicates detection techniques. It has also become extremely difficult to decide on the trustworthiness of such web documents when different versions/formats of the same content exist. Hence, the needs to bring in

semantics say meaningful comparison in near-duplicates detection with the help of the 6W factors – Who *(has authored a document),* What *(is the content of the document),* When *(it has been made available),* Where *(it is been available),* Why *(the purpose of the document),* How *(In what format it has been published/how it has been maintained).* This information can also be useful in calculating the trustworthiness of each document. A quantitative measure of how reliable that any arbitrary data is could be determined from the provenance information. This information can be useful in representative elimination during near-duplicate detection process and to calculate the trustworthiness of each document. The existing approaches of near-duplicates detection and elimination do not give much importance to the trustworthiness aspect of the content in documents retrieved through web search. Thus, Provenance based factors [2]. may be used for near-duplicates detection and elimination which provides the user with the most trustworthy results.

C. Lemmatization:-

Lemmatization refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma* . However, Tokenization is a fundamental step in processing textual data preceding the tasks of information retrieval, text mining, and natural language processing. Tokenization is a language dependent approach, including normalization, stop words removal, lemmatization and stemming. Both stemming and lemmatization share a common goal of reducing a word to its base. However, lemmatization is more robust than stemming as it often involves usage of vocabulary and morphological analysis, as opposed to simply removing the suffix of the word.

2. RELATED WORK

A. Stemming: *Stemming* refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. The most common algorithm for stemming English, is Porter's algorithm. Porter's algorithm consists of 5 phases of word reductions, applied sequentially. Within each phase there are various conventions to select rules, such as selecting the rule from each rule group that applies to the longest suffix. In the first phase, this convention is used with the following rule group:

(F)	Rule			Example		
	SSES	→	SS	caresses	→	caress
	IES	→	I	ponies	→	poni
	SS	→	SS	caress	→	caress
	S	→		cats	→	cat

Fig.1 Rule with Examples for Porter's Stemming

Other than , the Porter's Stemming algorithm , there are others stemming algorithms too.

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Fig.2 A comparison of 3 stemming algorithms on sample text

A. Comparison of Stemming and Lemmatization:

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance: am,are,is ⇒ be car, cars, car's, cars' ⇒ car

The difference between the two lies here: If confronted with the token *saw*, stemming might return just *s*, whereas lemmatization would attempt to return either *see* or *saw* depending on whether the use of the token was as a verb or a noun. The two may also differ in that stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma. Linguistic processing for stemming or lemmatization is often done by an additional plug-in component to the indexing process, and a number of such components exist, both commercial and open-source.

C. Semantic Approaches to detect near-duplicates: A method on plagiarism detection using fuzzy semantic-based[4] string similarity approach was proposed. The algorithm was developed through four main stages. First is pre-processing which includes tokenization, stemming and stop words removing. Second is retrieving a list of candidate documents for each suspicious document using shingling and Jaccard coefficient. Suspicious documents are then compared sentence-wise with the associated candidate documents. This stage entails the computation of fuzzy degree of similarity that ranges between two edges: 0 for completely different sentences and 1 for exactly identical sentences. Two sentences are marked as similar (i.e. plagiarized) if they gain a fuzzy similarity score above a certain threshold. The last step

is post-processing hereby consecutive sentences are joined to form single paragraphs/sections.

Recognizing that two Semantic Web documents[5] or graphs similar, and characterizing their differences is useful in many tasks, including retrieval, updating, version control and knowledge base editing. A number of text based similarity metrics are discussed as in that characterize the relation between Semantic Web graphs and evaluate metrics for three specific cases of similarity that have been identified: similarity in classes and properties used while differing only in literal content, difference only in base-URI, and versioning relationship.

3. PROPOSED WORK

In Web Search Optimization based on Web Provenance using Fuzzy Equivalence Relation in Web Document Clustering , the architecture comprises of the following components: (i) Data collection (ii) Pre-sprocessing(Tokenization, Stemming, Stop-Word Removal) (iii) Document Term Matrix(DTM) construction (iv) Provenance Matrix(PM) Construction (v) Database (vi) Singular Value Decomposition (vii) Document Comparison (viii) Fuzzy Clustering Equivalence method (ix) Filtering (x) Re-ranking based on trustworthiness values.

My Proposed Work is to refine the pre-processing concept by replacing stemming with Lemmatization in Web Search Optimization based on Provenance using Fuzzy Equivalence Relation in Web Document Clustering. This will provide better accuracy from the earlier concept and give much more refined results in the form of near-duplicates. Actually, 'lemmatization' can be used which is more robust than stemming as it often involves usage of vocabulary and morphological analysis, as opposed to simply removing the suffix of the word.

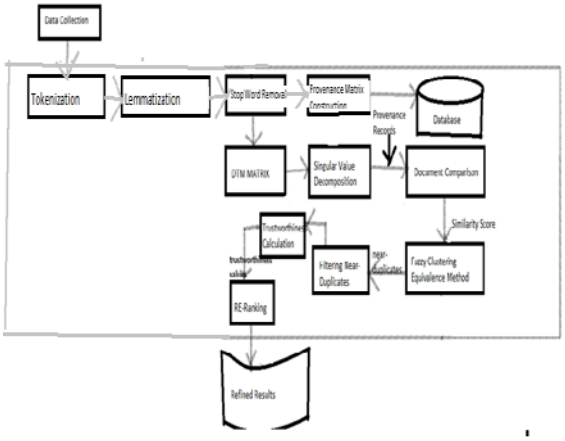

Figure 3: An Enhanced Approach to Optimize Web Search based on Provenance using Fuzzy Equivalence relation by Lemmatization.

4. CONCLUSION AND FUTURE WORK

In this paper, I have proposed a much more efficient way to optimize the web search for detecting and eliminating near-duplicates using provenance method in which fuzzy clustering equivalence method is used by refining the pre-processing concept by using lemmatization instead of stemming and this will give much better accuracy than the earlier method of Provenance having fuzzy clustering equivalence method. In future work, we can also focus on provenance matrix not using 'Why' and 'How' measure to detect near-duplicates or we can also work on clustering techniques to get highly similar document so that we get a much relevant document instead of redundant one.

ACKNOWLEDGMENT

Thanks to the authorities of their working organization and management for their support and encouragement to pursue research in the chosen field of study.

REFERENCES

[1] Tanvi Gupta, Optimizing Web Search Based on Web Provenance using Fuzzy Equivalence Relation in Web Document Clustering, IJESS Volume1Issue2 ISSN: 2249-9482, 2011.

[2] Y. Syed Mudhasir, J. Deepika, S. Sendhilkumar, G. S. Mahalakshmi, Near-Duplicates Detection and Elimination Based on Web Provenance for Effective Web Search in International Journal on Internet and Distributed Computing Systems. Vol: 1 No: 1, 2011

[3] Anjali B. Raut and G. R. Bamnote Web Document Clustering Using Fuzzy Equivalence Relations in Journal of Emerging Trends in Computing and Information Sciences, Volume 2 Special Issues (2010-11) CIS

[4] Salha Alzahrani and Naomie Salim, Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection, 2010.

[5] Krishnamurthy Koduvayur Viswanathan and Tim Finin, Text Based Similarity Metrics and Delta for Semantic Web Graphs, pp: 17-20, 2010.

[6] Joël Plisson, Nada Lavrac, Dunja Mladenic, A Rule based Approach to Word Lemmatization.

[7] Eiman Al-Shammari, Jessica Lin, A Novel Arabic Lemmatization Algorithm, *AND'08*, July 24, 2008,Singapore.(Copyright © 2008 ACM)

[8] Ilia Smirnov, Overview of Stemming Algorithms, 2008.