Research Journal of Applied Sciences, Engineering and Technology 9(5): 359-364, 2015
DOI:10.19026/rjaset.9.1414
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 18, 2014 Accepted: October 01, 2014 Published: February 15, 2015

Research Article

Efficient Clustering of Web Search Results Using Enhanced Lingo Algorithm

[1]M. Manikantan and [2]S. Duraisamy
[1]Department of Computer Applications, Kumaraguru College of Technology,
[2]Department of Computer Applications, Sri Krishna College of Engineering and Technology,
Coimbatore, India

Abstract: Web query optimization is the focus of recent research and development efforts. To fetch the required information, the users are using search engines and sometimes through the website interfaces. One approach is search engine optimization which is used by the website developers to popularize their website through the search engine results. Clustering is a main task of explorative data mining process and a common technique for grouping the web search results into a different category based on the specific web contents. A clustering search engine called Lingo used only snippets to cluster the documents. Though this method takes less time to cluster the documents, it could not be able to produce the clusters of good quality. This study focuses on clustering all documents using by applying semantic similarity between words and then by applying modified lingo algorithm in less time and produce good quality.

Keywords: Clustering algorithm, lingo algorithm, search engine optimization, semantic web, snippet, web query optimization

INTRODUCTION

Today, every individual are one way or other using internet search directly or indirectly. Comparing the internet usage in last two decades and now we can understand this truth. To fetch required information the users are using search engines or sometimes through the website interfaces. Even though there are many search engines today, Google™ tops on all other peers. One can give his or her web query and get innumerable number of results page after page. User is sometimes tired in getting the information. Also not all result pages are analyzed by the user, for example, if the result gives say 100 pages, the user normally will surf the first 5 or 10 pages only and also other pages are irrelevant to his expected result. There are more people working on this problem to get as relevant results as possible. One approach is search engine optimization which is used by the website developers to popularize their website through the search engine results.

Search engines do a sort of text mining to find the words repeated in each pages of the website through something called frequency of the words when it is mined through text mining. The each page is approximately linked by the keywords and also even the websites are also categorized by the keywords as like web pages. For example, the websites are categorized as a sports website or academic website, or banking website etc. Though such information can be obtained through the home page by analyzing html tags for <title>, <Meta> etc, but by using text mining, deeper categorization can be done.

This study combines web query optimization through text mining by categorizing the results and we get clustering of web query results so that instead of getting a big long result pages we get clusters of user interested clusters under which pages of the specific cluster contains only relevant information. The attributes of interest are relevance of the results, summary of results by clustering, snippet tolerance and speed. Here instead of clicking the result hyperlink to know the entire page, we use snippet, which is small paragraph like information of the web query is displayed in the list of results in each clusters pages. Hence the user can go through the snippet and if it matches his expected result he can further click it and get the entire page or otherwise he can just skip it and this is the main advantage of this study. This project focuses on clustering all documents using by applying semantic similarity between words first and then by applying Lingo algorithm.

LITERATURE SURVEY

With the developments of Internet, the Internet age in the society has emerged inevitably. Web search engines have become an important part of Internet usage (Kantabutra, 2001). Most of the search engines

Corresponding Author: M. Manikantan, Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, India

provide millions of search results. The results maybe scattered which requires the user to scan through the results to obtain what he wants. Also existing search engines drawbacks are, Poor precision-List of retrieved documents contains high percentage of irrelevant documents. Poor recall-This might lead to situation wherein not even one of the top ten sites listed would be of subject the user expects.

Search engines results are mainly web pages or just documents in html or other web forms. While relating the search words with search result pages, non-informative words need to be removed since commonly used document processing technique in text classification and text filtering minimizes the redundancy of computation. Also clustering techniques need to be used by clustering algorithm with reduced storage space used and time for computing. These Non-informative words are often defined by "stop word list" which typically consists of about 400 terms including articles, prepositions, conjunctions and certain high frequency words like verbs, adverbs and adjectives. Now let us consider in what way clustering can be used in categorizing the result pages of a search engine. Analyzing query logs has a broad impact in different applications for Web searching such as Web availability, document and index caching and Web crawling (Correia-Saravia *et al*., 2001). It also proposed a clustering framework that allows one to find groups of semantically related queries. Our experiments show that the bias reduction technique proposed improves the quality of the clusters found. The results also provide evidence that our ranking algorithm improves the retrieval precision of the search engine and that our query recommender algorithm has good precision in the sense that it returns relevant queries to the input query. The search for certain groups of queries capturing common sets of preferences and information needs has been a recent trend in query log analysis (Beeferman and Berger, 2000; Wen *et al*., 2001; Zhang and Dong, 2002). It also proposes a query clustering technique based on common clicked URLs (Beeferman and Berger, 2000). And Wen *et al*. (2001) proposes clustering similar queries to recommend URLs to frequently asked queries of a search engine. They use four notions of query distance:

- Based on keywords or phrases of the query
- Based on string matching of keywords
- Based on common clicked URLs
- Based on the distance of the clicked documents in some predefined hierarchy

Clustering provided an organized way to manage a search engine. With the huge growth of web pages, it is difficult for users to find the relevant document of their interests easily. By applying clustering, data is collected from websites pages with features like their title length, number of keywords, URL length, number of back links, in links. Based on these parameters clusters are made to derive the conclusion (Minky and Nisha, 2013). Also Page ranking algorithms based on links (e.g., Most-Cited (Zhexue, 1998), Page Rank (Brin and Page, 1998) and hits (Kleinberg, 1998a, b) have been described and classified by Lawrence and Giles (1999) are used for clustering.

Cluster analysis or clustering is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main task of explorative data mining and a common technique for statistical data used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics. "Clustering based on k-means" that it is closely related to a number of other clustering and location problems which include the Euclidean k-medians which minimize the sum of distances to the nearest center and the geometric k-center problem, which aimed to minimize the maximum distance from every point to its closest center (Hongwei, 2010). K-Means clustering is a very popular algorithm to find the clusters in a dataset by iterative computations. It has the advantage of simple implementation and finding at least local optimal clustering. K-Means algorithm is employed to find the clustering in dataset (Wang and OuYang, 2010).

The algorithm (Kantabutra, 2001) is composed of the following steps:

- Initialize k cluster centers to be seed points. (These centers can be randomly produced or use other ways to generate).
- For each sample, find the nearest cluster center, put the sample in this cluster and recomputed centers of the altered cluster (Repeat n times).
- Exam all samples again and put each one in the cluster identified with the nearest center (don't recomputed any cluster centers). If members of each cluster haven't been changed, stop. If changed, go to step 2.

Using Weka tool, one can solve clustering of data by using data file with option which is in arff or csv format (Wagsta and Cardie, 2000; Zhao and Karypis, 2004). Also Ricardo *et al*. (2007) presents a framework for clustering Web search engine queries with the aim to identify groups of queries used to search for similar information on the Web. This framework is based on a novel term vector model of queries that integrates user selections and the content of selected documents extracted from the logs of a search engine.

PROPOSED METHODOLOGY

The proposed method for web search results clustering comprises by the four different phases,

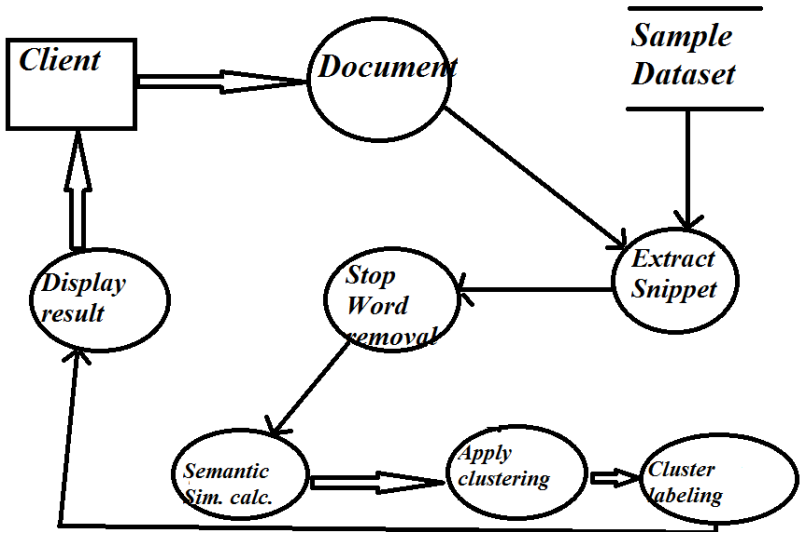

Fig. 1: Overall process for proposed method

named 1. Term extraction, 2. Stop word removal, 3. Semantic similarity calculation, 4. Applying clustering algorithm. The overall process for proposed method is shown in Fig. 1.

The enhanced lingo clustering algorithm:

1: D <- Input documents (or Snippets)
{**Step 1: Preprocessing**}
2: for all d Є D do
3: Perform text segmentation of d;
4: if language of d recognized then
5: mark stop-words in d;
6: end if
7: end for

The above step searches the entire documentation for stop words and marks it for removal.

{**Step 2: Semantic similarity calculation**}
8: T <- Set of Terms
9: for each pair of terms ti Є T and tj Є T do
10: Find semantic similarity SSij using wordnet
11: if SSij> = Semantic Similarity Threshold
12: Replace the term tj with ti
13: end if
14: end for

The second step as above is used to check the semantic similarity of key terms with search words with a semantic similarity threshold and similar words are identified with unique root key word.

{**Step 3: Frequent phrase extraction**}
15: Concatenate all documents;
16: Pc <- discover complete phrases;
17: Pf <- p: {p Є Pc ^ frequency (p) >Term Frequency Threshold};

The step 3 is used to check whether the frequency of the given key word is above frequency threshold only those words that satisfy this condition get into the corresponding cluster.

{**Step 4: Cluster label induction**}
18: A <- term-document matrix of terms not marked as stop-words and with frequency higher than the Term Frequency Threshold
19: ∑, U, V <- SVD (A); {Product of SVD decomposition of A}
20: k <- 0; {Start with zero clusters}
21: n <- rank (A)
22: repeat
23: k<-k+1
24: $q \leftarrow (\sum_{i=1}^{k} \sum_{i=1}^{n} ii)/(\sum_{i=1}^{n} \sum_{i=1}^{n} ii)$
25: until q<Candidate Label Threshold
26: P<- phrase matrix of Pf
27: Calculate Mij = abs$[(U_k^T P)_{ij}]$

Here in step 4 SVD decomposition of the resultant cluster words are used to match the given phrase in a phrase matrix.

{**Step 5: Cluster content discovery**}
28: for all L Є Cluster Label Candidates do
29: create cluster C described with L
30: add to C all documents whose similarity to C exceeds the Snippet Assignment Threshold
31: end for

Finally, in step 5 clusters are formed with a conditional parameter called Snippet Assignment Threshold (SAT).

In our algorithm, the cluster labels are automatically assigned by using processing which are Term extraction and Stop word removal with stemming. Whereas in Lingo algorithm the primary aim of the preprocessing phase is to remove from the input documents all characters and terms that can possibly affect the quality of group descriptions.

In Lingo, Semantic similarity calculation, a semantic similarity threshold is used which was set to 0.7. The terms exceed this threshold are considered for further steps that is, the words exceeding the threshold value are replaced with its equivalent semantic word. Whereas in the proposed system, this second step is used to check the semantic similarity of key terms with search words with a semantic similarity threshold and similar words are identified with unique root key word which is identified after Term extraction and Stop word removal with stemming. In Lingo, Frequency phrase extraction works in two steps. In the first step, the right and left complete phrases are discovered and in the second step, they are combined into a set of complete phrases. In the final step of the feature extraction phrase, terms and phrases that exceed the Term Frequency Threshold is chosen. In our proposed system it is used to check whether the frequency of the given key word obtained in first two steps is above frequency threshold and only those words that satisfy this condition get into the corresponding cluster. Step 4 and 5 of Lingo algorithm achieved by considering the following sub steps, 1.Term-document matrix building 2. Abstract concept discovery 3. Phrase matching, 4. Label pruning and evaluation and also the input snippets are assigned to the cluster labels induced in the previous phase. These steps are combined into one step called clustering algorithm in the proposed system that assigns snippets to the cluster labels with sample contents. Also SVD decomposition of the resultant cluster words is used to match the given phrase in a phrase matrix in step 4 of the proposed system.

Lingo achieves impressing empirical results, but the work on the algorithm is obviously not finished. Cluster label pruning phase could be improved by adding elements of linguistic recognition of nonsensical phrases. Topic separation phase currently requires computationally expensive algebraic transformations. It is tempting to find a method of inducing hierarchical relationships between topics. Finally, a more elaborate evaluation technique will be necessary to establish weak points in the algorithm.

The proposed system is an incremental approach with small memory footprint which is of great importance for our algorithm with more scalability and also it solves weak points of Lingo algorithm with first two detailed steps of our proposed algorithm.

RESULTS AND DISCUSSION

For our experiment, academic domain is chooses for sample data set containing with web sites. Then our modified lingo clustering algorithm is applied to all the sample web site documents. The snippets are retrieved from the web documents and stop words are removed from the snippets as they do not contribute much to clustering and also it reduces the processing time greatly. The preprocessing module takes snippets from the sample dataset and removes stop words. The stop words are saved in a text file. After removing the stop words, the resultant words are stored in a table called "semantics." Then the semantic similarity is measured between the terms and they replaced if the similarity between them is greater than or equal to 0.7. The figures show the "semantics" before and after semantic measure calculation. The similarity measure is calculated using the Wu and Palmer method. It finds the similarity using the relationship between terms. Frequent phrases extracted from the snippets are stored in a table named "third word." All the two worded and three worded phrases are generated. Then from the phrases generated the frequent phrases which satisfy the threshold frequency are stored in a separate table named "freq_threshold." For a given query, the words are extracted from the <Title>, <META>, Snippets and links stored in the data base. Stop words are removed from the extracted words. Then stemming is performed and the unique keywords are identified. Based on this algorithm, the query "Best Schools in Chennai" was processed by this technique. The query results are automatically clustered by this algorithm instead of displaying the sequence order of page by page display of web search. The results are tabulated and given below in Table 1. Also the different types of user queries in the academic domain were examined by this method with different datasets. For this process, academic domain data sets were taken and both methods were implemented. All the query results are clustered by specific titles based on their content with the processing time and total number of web sites. The final results are compared in both methods and represented in Table 2. Figure 2 shows the processing time for clustering the web sites based on the user query of both methods. The y axis shows the time taken for query processing and clustering the web sites (in milliseconds) and x axis shows the number of web sites (domain size) participated in the query process.

Table 1: Cluster results for the user query of "Best schools in Chennai"

Cluster no.	Cluster name	No. of sites under the cluster
1	"CBSE School"	25
2	"State Board School"	35
3	"Montessori School"	10
4	"Preschool"	12
5	"Play School"	7

Table 2: Comparison results for processing time and cluster count

	Web query processing time and clusters count in academic domain			
No. of sites	Modified lingo algorithm	No. of clusters	Lingo algorithm	No. of clusters
100 sites	6200 ms	5	6700 ms	4
200 sites	6500 ms	6	7200 ms	5
300 sites	6800 ms	6	7700 ms	5
400 sites	7200 ms	7	8300 ms	6
500 sites	7600 ms	8	8900 ms	6

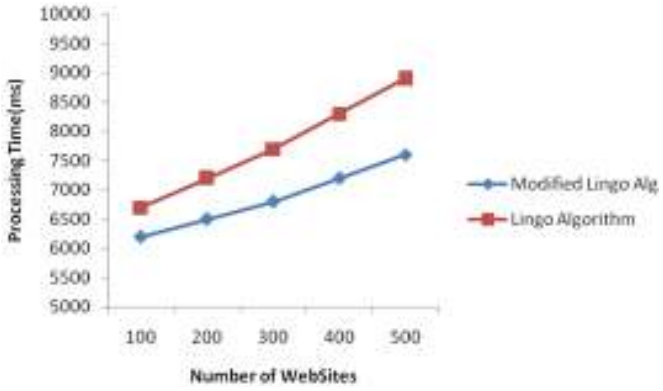

Fig. 2: Processing time comparison for both algorithms

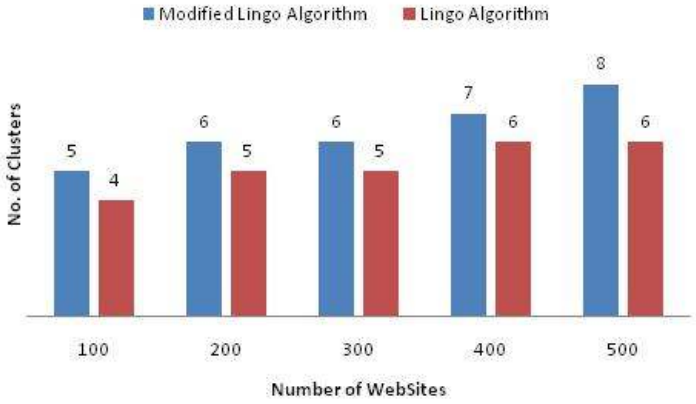

Fig. 3: Clusters count comparison for modified lingo vs. lingo algorithm

The graph in Fig. 3 shows the clusters count based on the user query of both methods. The y axis shows the number of clusters after the query processed and x axis shows the number of web sites (domain size) participated in the query process.

CONCLUSION AND RECOMMENDATIONS

The proposed system is used to searches the web documents and clusters the results. Clustering based on snippets rather than entire documents are used for clustering. This reduces the processing time for searching and grouping the web results with more number of clusters. Semantic similarity is calculated between the terms in the documents and then clustered by the similarity mean. The existing system makes use of entire keywords which has a higher processing time. This system provides the user with clustered results which reduces the search complexity for the user. This system can be further enhanced by pre-indexing the web pages in a local cache which further reduces the time factor. The system can be made to remove redundancy in the clustered results by eliminating the exact copies of web pages. Extension of this study is possible by comparing the efficiency of other intelligent clustering techniques with decreasing the processing time and increasing the cluster range.

REFERENCES

Beeferman, D. and A. Berger, 2000. Agglomerative clustering of a search engine query log. Proceeding of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, pp: 407-416.

Brin, S. and L. Page, 1998. The anatomy of a large-scale hyper textual web search engine. Comput. Netw., 30(17): 107-117.

Correia-Saravia, P., E. Silva de Moura, N. Ziviani, W. Meira, R. Fonseca and B. Ribeiro-Neto, 2001. Rank-preserving two-level caching for scalable search engines. Proceeding of the 24th International ACM Conference on Research and Development in Information Retrieval. New Orleans, LA, pp: 51-58.

Hongwei, Y., 2010. A document clustering algorithm for web search engine retrieval system. Proceeding of the International Conference on e-Education, e-Business, e-Management and e-Learning, pp: 383-386.

Kantabutra, S., 2001. Efficient representation of cluster structure in large data sets. Ph.D. Thesis, Tufts University, Medford, MA.

Kleinberg, J., 1998a. Authoritative sources in a hyperlinked environment. Proceeding of the ACM-SIAM Symposium on Discrete Algorithms.

Kleinberg, J.M., 1998b. Authoritative sources in a hyperlinked environment. J. ACM, 46(5): 604-632.

Lawrence, S. and C.L. Giles, 1999. Searching the web: General and scientific information access. IEEE Commun. Mag., 37(1): 116-122.

Minky, J. and K. Nisha, 2013. K-means clustering technique on search engine dataset using data mining tool. Int. J. Inform. Comput. Technol., 3(6): 505-510.

Ricardo, B.Y., H. Carlos and M. Marcelo, 2007. Improving search engines by query clustering. J. Am. Soc. Inform. Sci. Technol., 58(12): 1793-1804.

Wagsta, K. and C. Cardie, 2000. Clustering with instance-level constraints. Proceeding of the 17th International Conference on Machine Learning. Palo Alto, Morgan Kaufmann, CA, pp: 1103-1110.

Wang, J. and Z.Z. OuYang, 2010. The research of K-means clustering algorithm based on association rules. Proceeding of the International Conference on Challenges in Environmental Science and Computer Engineering, pp: 285-286.

Wen, J., J. Nie and H. Zhang, 2001. Clustering user queries of a search engine. Proceeding of the International Conference on World Wide Web. Hong Kong, China, pp: 162-168.

Zhang, D. and Y. Dong, 2002. A novel web usage mining approach for search engines. Comput. Netw., 39(3): 303-310.

Zhao, Y. and G. Karypis, 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. Mach. Learn., 55(3).

Zhexue, H., 1998. A fast clustering algorithm to cluster very large categorical data sets in data mining. Cooperative Research Centre for Advanced Computational Systems (ACSys) Established under the Australian Government's Cooperative Research Centres Program.