News Retrieval Based On Latent Semantic Index And Clustering

Prerna, Rajesh Singh, Pawan Bhadana

***Abstract*:-** Web is a collection of heterogeneous as well as unstructured set of news articles. This paper presents a novel approach to retrieve relevant news articles from heterogeneous and unstructured collection of articles. Efficient retrieval requires analysis of news articles based on keyword. Two problems that occur in the analysis of news articles are synonymy and polysemy. In this paper, we present a News Retrieval approach based on Latent Semantic Index (LSI) and Clustering. It includes projection of keyword-news article matrix into small spaces called clusters. After that, clustering approach is used to group relevant articles into clusters.

***Keywords*: -** Extraction, Algorithm, Mining

I. INTRODUCTION

Search engines are used to retrieve relevant news articles from news collection. This retrieval of efficient news articles depends on the how relevant news articles accessed and satisfaction of users with the retrieved articles. Latent Semantic Indexing (LSI) is introduced for news retrieval combining with a clustering algorithm. We know that the problem of synonymy and polysemy arises in retrival process. Synonymy represent differenty words but meaning of them is identical or similar. Polysemy represents words with multiple meaning. Earlier, Vector Space Model(VSM) was introduced for retrieval based on keyword searches. Word in articles are represented with mathematical vectors that are one dimensional arrays. Importance of keywords in articles are evaluated with keyword frequency(tf) and inverse document frequency(idf). This scheme (tf-idf) is designed to consider the discriminative power of keyword within an article and over articles. News Retrieval based on Latent Semantic Index and Clustering includes searching news articles within databases which could either be relational stand-alone databases or hyper textually-networked databases like the World Wide Web. It includes locating from a large news articles collection; those articles that fulfill a specified information need by creating indexes on the basis of concept and then forming clusters.

- Prerna, Rajesh Singh, Pawan Bhadana
- Student M.Tech. (CSE), B. S. Anangpuria Institute of Technology and Management, Faridabad,India Email: prernahooda@gmail.com
- Assistant Professor, B. S. Anangpuria Institute of Technology and Management, Faridabad,India Email: rajeshsingh22@gmail.com
- 3HOD CSE, B. S. Anangpuria Institute of Technology and Management, Faridabad,India
  Email: Pawanbhadana79@gmail.com

In the following sections we will discuss News Retrieval based on Latent Semantic Index and Clustering for relevant news retrieval. The rest of this paper is organized as follows. Section 2 briefly introduces the related approach of news retrieval. In section 3, we introduce our novel method of News Retrieval based on Latent Semantic Index and Clustering. Section 4 summarizes the paper and outlines some interesting directions for future research.

II. RELATED WORK

There has been surge of interest in document clustering after Lee and Sung's [3] update rules for NMF proved to perform better than Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD). Many researchers are approaching with efficient algorithms and comprehensive comparison of the existing algorithms. Latent Semantic Indexing (LSI) is a novel Information Retrieval technique that was designed to address the deficiencies of the classic VSM model[5]. Document clustering can loosely be defined as “clustering of documents”.[2] Clustering is a process of understanding the similarity and/or dissimilarity between the given objects and thus, dividing them into meaningful subgroups sharing common characteristic.[11] Good clusters are those in which the members inside the cluster have quite a deal of similar characteristics. Various mathematical models have been proposed to represent Information Retrieval systems and procedures; the boolean model, the probabilistic model, the vector space model etc. Of the above models, the vector space model has been quite widely used model[9]. Well-known search engines like Google and Yahoo also extract information from web pages and categorize them according to topic. Bar-Yossef and Rajagopalan in [5] Ho present methods to extract informative information from web page tables. Ramaswamy et al. in [3] also presented the same method. An approach to detect content structure on web pages based on visual representation was presented by Cai et al. [10]. Embley et al. [15] present heuristics for extracting records from web pages which is a domain specific approach. Traditional methods in document clustering use words as measure to find similarity between documents. [7]These words are assumed to be mutually independent

which in real application may not be the case. Traditional VSI uses words to describe the documents but in reality the concepts/semantics/features/topics are what describe the documents. The extraction of these features from the documents in Feature Extraction. The extracted features hold the most important idea/concept pertaining to the documents.

III. PROPOSED WORK

News Retrieval based on Latent Semantic Index and Clustering is based on projecting keyword-news article matrix into small clusters. The dimensional reduction of a matrix is accomplished using singular value decomposition which decomposes the keyword-news article matrix into three matrices that are article eigen vector matrix, eigen value matrix and term eigen vector matrix. Original article matrix is obtained by multiplying these three matrices with only high eigen values. Here, we consider the fact that orthogonal characteristic of matrix, keywords in matrix have little relations with keywords in other matrix, but have high relation with keywords in the same matrix. This characteristic deals with synonymy of keyword. Latent Semantic Index allows articles to be retrieved based on keyword matching even though they are not indexed by the query index terms. After Latent Semantic Indexing, we apply a clustering algorithm that groups similar articles into clusters. Each cluster is defined by a centroid that is placed far from each other as possible. The article is added to the cluster which the centroid is similar to. When all articles are added to the cluster, the centroids are then recalculated.. with nerw centroids, new clusters are created. Algorithm is applied that returns the number of clusters formed from new articles based on keywords that are occurring in the article. Fig 1 presents the algorithm.

1. Input: initial number of k clusters
2. repeat
3. {
4. n=|newsarticle|;
5. for j = 0 to n
6. set min=sim[N_j,C_0];
7. t=0;
8. for i= 1 to k
9. if sim[N_j.C_i]<min then
10. min= sim[N_j.C_i]
11. t=i;
12. end if
13. end for
14. C_t=N_j
15. Enf for
16. For i= 1 to k do
17. Recalculate centroid positions for C_i
18. End for
19. Until centroids do not move
20. Return k clusters of news articles

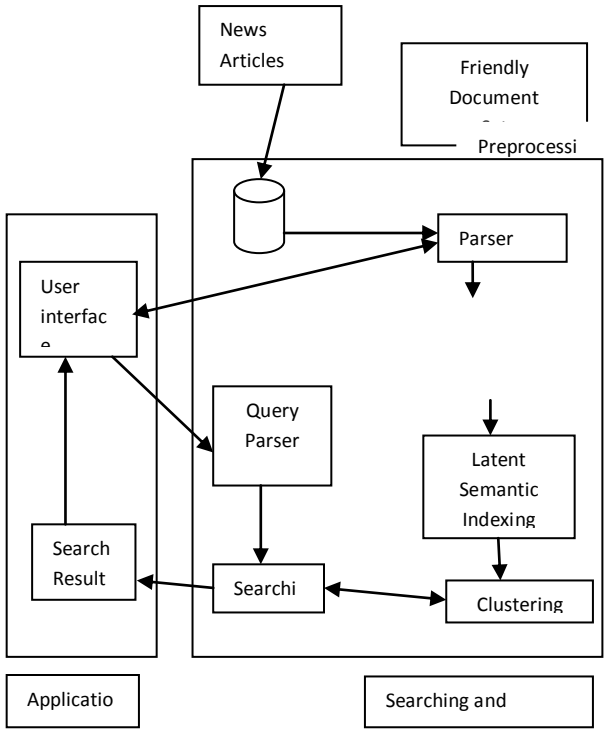

Fig 2 News Retrieval Based on Latent Semantic Index and Clustering Architecture

IV. EXPERIMENT AND EVALUATION

Latent Semantic Index and Clustering System architecture shown in Fig 2 is built to implement indexing and clustering for efficient news retrieval. The system architecture includes three main modules: Preprocessing, Indexing and Clustering and User Interface. Preprocessing consists of set of news articles. A parser is used that parses and split news articles into a set that can be indexed. Indexing and clustering module is used to index articles and parse queries. Indexing of articles done by using Latent Semantic Index forms set of indexed articles. Clustering algorithm reads indexed articles. This algorithm initially guesses the allocation of centroid vectors and forms the initial clusters. Indexed articles are added to the clusters with highest similarity until the centroid location moves. These clusters are returned. User interface is built that takes user query as input that is passed to the Clustering and Indexing module. Here searching is done with clusters that are were initially formed from Indexed articles. The relevant articles are returned to the user interface.

V. CONCLUSION

This paper presents News Retrieval based on Latent Semantic Index and Clustering approach for relevant news retrieval. We proposed this system to improve retrieval process through indexing and then cluster formation. There are still some issuestaht nned to address. The intitial allocation of centroid vectors is completely different from the clusters built later. If a cluster doesnot get any indexed articles,it remains empty. Regardless of the the above

issues, our News Retrieval based on Latent Semantic Index and Clustering approach can be used to retieve relevant news articles from a set of news articles.. our future work includes improving the above issues.

REFERENCES

[1]. Cutting, D, Karger, D, Pederson, J & Tukey, J (1992). Scatter/gather: A clusterbased approach to browsing large document collections. In Proceedings of ACM SIGIR.

[2]. Guduru, N (2006). Text mining with support vector machines and nonnegative matrix factorization algorithm. *Masters Thesis. University of Rhode Island, CS Dept.*

[3]. Xu, W, Liu, X & Gong, Y (2003). Document clustering based on nonnegative matrix factorization. Proceedings of ACM SIGIR, pages 267–273.

[4]. Dhillon, SI & Modha, DS (2001). Concept decompositions for large sparse text data using clustering.

[5]. Landauer, T, Foltz, PW & Laham, D(1998). Introduction to Latent Semantic Analysis.. Discourse Processes 25: pages 259–284.

[6]. Michels, S (July 5, 2007). Problem Solving on LargeScale Clusters, Lecture 4.

[7]. Dean, J & Ghemawat, J (December 2004). MapReduce: Simplified Data Processing on Large Clusters. In the Proceedings of the 6th Symp. on Operating Systems Design and Implementation.

[8]. Tropp, J., *An Alternating minimization algorithm for non-negative matrix approximation.*

[9]. Lee, D., Seung, H., *Learning the Parts of Objects by Non-negative matrix factorization in Nature* (1999).

[10]. Tong, S, Koller, D., *Support Vector Machine Active Learning with Applications to Text Classification.*

[11]. Lee, D., Seung, H.S., *Learning the Parts of Objects by Non-negative matrix factorization in Nature* (1999).

[12]. Florian Beil, Martin Ester, and Xiaowei Xu. “Frequent Term-Based Text Clustering”, In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining New York, NY, USA.

[13]. Raymond Kosala and Hendrik Blockeel, “Web Mining Research: A survey”, SIGKDD Exploration, Vol.2 issue 1, July 2000, pp- 1-15.

[14]. Aura Conci., Everest Mathias M. M. Castro “Image Mining By Color Content “

[15]. Zhang Ji, Wynne Hsu, Mong Li Lee, “Image Mining: Issues, Frameworks and Techniques”, in Proc. of the 2nd International Workshop on Multimedia Data Mining (MDM/KDD'2001), San Francisco, CA, USA, 2001, pp. 13-20.