Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: **1992-8645** www.jatit.org E-ISSN: **1817-3195**

IMPROVING INFORMATION RETRIEVAL USING DOCUMENT CLUSTERS AND SEMANTIC SYNONYM EXTRACTION

[1]G.BHARATHI, [2] D.VENKATESAN
[1] School of Computing, SASTRA University, Tamil Nadu, India.
[2]Assistant Professor, School of Computing, SASTRA University, Tamil Nadu, India.
E-mail: [1]bharathi_gin@yahoo.com, [2] venkatgowri@cse.sastra.edu

ABSTRACT

Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems and as an efficient way of finding the nearest neighbors of a document. More recently, clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to a user's query. This paper presents a new semantic synonym based correlation indexing method in which documents are clustered based on nearest neighbors from the document collection and then further refined by semantically relating the query term with the retrieved documents by making use of a thesaurus or ontology model to improve the performance of Information Retrieval System (IRS) by increasing the number of relevant documents retrieved. Results show that the proposed method achieves significant improvement than the existing methods and may generate the more relevant document in the top rank.

Keywords: *Document Cluster, Information Retrieval, Semantic Similarity, Correlation Preserving Indexing (CPI).*

1. INTRODUCTION

Data mining is a technique to get the pattern from hidden information. This technique is to find and describe structural patterns in data collection as a tool for helping to explain that data and make predictions from it. Generally, data mining tasks are divided into two major categories: predictive tasks which aim to predict the value of a particular attribute based on the values of other attributes and another one is descriptive tasks which aim to derive patterns (correlations, trends, clusters, trajectories, and anomalies) [1]. Clustering is a method to organize automatically a large data collection by partition a set data, so the objects in the same cluster are more similar to one another than to objects in other clusters. Document clustering is a fundamental task in text mining that is concerned with grouping documents into clusters according to their similarity. In the field of Information Retrieval (IR), Information users could encounter the following problem when interacting with the document collection:

Finding relevant information: People either browse or use the search service when they want to find specific information on the Web. When a user uses search service he or she usually inputs a simple keyword query and the query response is the list of pages ranked based on their similarity to the query. However today's search tools have the following problems. The first problem is low precision, which is due to the irrelevance of many of the search results. This results in a difficulty finding the relevant information. The second problem is low recall, which is due to the inability to index all the information available on the Web. This results in a difficulty finding the unindexed information that is relevant.

Document clustering is used to automatically group the documents that belong to the same topic in order to provide user's browsing of retrieval results [2]. Some experimental evidences show that IR application can benefit from the use of document clustering [3]. Document clustering has always been used as a tool to improve the performance of retrieval and navigating large data. Most of document clustering algorithms use the vector space model (VSM) [4], [5] alone for document representation. VSM represents documents as vectors in the space of terms and uses the cosine similarity between document vectors to estimate their similarity. VSM, however, ignores any semantic relations between terms.

Sometimes by matching only query terms to document, it is not possible to retrieve relevant documents. This motivates the proposed work to prepare document clusters based on nearest neighbors cluster similarity so that it will not only retrieve the documents which contain query terms but also retrieve those documents which are similar to retrieved document. First the n nearest neighbors of all points are found. If two data points are similar enough, they are considered as neighbors of each other. Every data point can have a set of neighbors in the data set for a certain similarity threshold. The documents are ranked only on the basis of the term frequency. This motivates the proposed model to give importance to the information content of the document under consideration. Thesaurus or Ontology as background Knowledge has been applied to various text mining problems but very few attempts have been made to utilize it for document clustering. Here, in this paper, we utilize the online encyclopedia Conservapedia, to retrieve the synonyms of the query term so that from the retrieved documents of the dataset the correlated semantic terms of the specified query term are identified and finally more similar documents are ranked based on semantic correlation similarity. This improves the accuracy of the retrieved relevant documents without much increasing time.

The rest of this paper is organized as follows: Next Section reviews the related works about document clustering, Section 3 briefly describes the proposed methodology, Section 4 discusses the experimental setup and performance measures and Section 5 concludes the paper and discusses about the future work.

2. RELATED WORK

Some work has been proposed on using Latent Semantic Indexing (LSI) for document clustering [6]. However, most of dimension reduction techniques are computationally very expensive for large data sets, and they suffer from the problem of determining the intrinsic dimensionality of the data. Other models for document representation are based on analyzing the semantics of terms using a lexical database, such as WordNet [7]. Hotho et al [8], for instance, proposed an approach in which terms are augmented or replaced by vectors based on related concepts from WordNet. Document clustering algorithms are then applied to these concept vectors. Recent work on document clustering constructs concept vectors using Wikipedia [9]. These methods, however, are computationally complex and produce high-dimensional concept vectors. Some related work for document clustering is based on explicitly grouping related terms first, and then grouping documents into clusters based on term clusters [10].Simultaneous clustering of terms and documents [11] is a related approach which is based on spectral partitioning of a bipartite graph of documents and terms. These methods, however, do not scale well to handle large data sets. Numerous documents clustering algorithms appear in the literature. The two most common techniques used for clustering documents are Hierarchical and Partitional (K-means) clustering techniques.

3. PROPOSED METHODOLOGY

In this paper, a method is proposed in which document clusters are formed from the document set by using term frequencies and term weighting of each document, finding their synonyms using Conservapedia and then applying K-means partitioning algorithm based on the concept of nearest neighbors[12]. This is done by considering the term frequencies and document frequencies and preparing a document summary for each cluster containing the distinct terms whose frequencies are high after preprocessing of the documents. This is done to measure the document relevant score. Documents are then semantically correlated using Conservapedia to find out the synonyms of the terms in the document summary and thereby clustering those documents that are semantically similar to each other as in Figure 1. After implementation of the above methodology it was found that proposed work is better than the existing methods.
The steps involved are as below:

1. Document preprocessing which includes case folding, parsing, removing stop words and stemming. This step is done to reduce high dimensionality of the data set and improve the computational time.

1.1) Filtering: Filtering removes special characters and punctuation from documents, which are not thought to hold any discriminative power under the vector model.

1.2) Tokenization: This step splits sentences into individual tokens, typically words.

1.3) Stemming: This is known as the process of reducing words to their base form, or stem. For example, the words "connected", "connection", "connections" are all reduced to the stem "connect". Porter's algorithm is the de facto standard stemming algorithm. A smaller number of distinct terms results in a saving of memory space and processing time.

1.4) Stop word removal: A stop word is defined as a term which is not thought to convey any meaning as a dimension in the vector space. Stop words are the most common words (e.g., "and", "or", "in") in a language, but they do not convey any significant information so they are stripped from the document set.

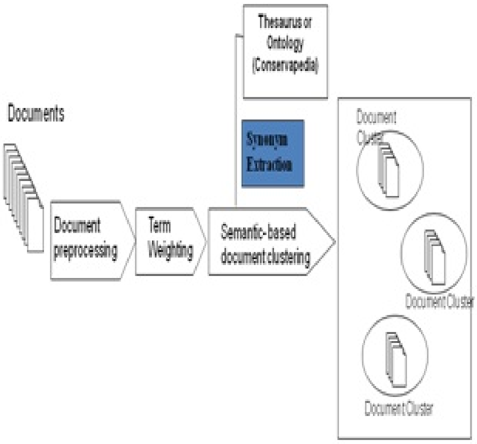

Figure 1. Document Clustering

1.5) Document Representation and Term weighting: The various clustering algorithms use the *tf-idf model* for term weighting.The full name of tf-idf is term frequency-inverse document frequency. In this model, each document D is consider to be represented by a feature of the form (d_1, d_2,…,d_n), where d_i is a word in D. The order of the d_i is based on the weight of each word. The formula below is the common weight calculating formula that is widely used, while in different approaches, the formula is not exactly the same. Some extra parameters may be added to optimize the whole clustering performance. Such as in [13], weight calculating formula in tf-idf:

$$w_{ij} = tf_{ij} \cdot idf_j$$

$$\text{where } idf_j = \log\left(\frac{n}{df_j}\right)$$

…................. (1)

In the above formula, each factor is explained below:

tf_{ij} is number of occurrences of the term t_j in the Web page P_i

idf_j is Inverse document frequency.

df_j is the number of Web pages in which term t_j occurs in the web document collection.

n is the total number of Web pages in the database.

From this formula, we can see that the weight of each term is based on its inverse document frequency (IDF) in the document collection and the occurrences of this term in the document..

2. Preparing a document summary of the terms with highest frequency for each document.
3. Semantic synonyms extraction for the terms in the document summary using Conservapedia.
4. Implementing K-means clustering algorithm by term-document vector as representative of the document collection and using Cosine similarity as replacement of Euclidean distance. For Clustering we have chosen Partitional clustering algorithm K-means. It has been recognized that Partitional algorithms are better suited for handling large document datasets than hierarchical ones, due to their relatively low computational requirements [14, 15].

The standard K-means algorithm works as follows. Given a set of data objects D and a pre-specified number of clusters k, k data objects are randomly selected to initialize k clusters, each one being the centroid of a cluster. The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid. Next, new centroids are recomputed for each cluster and in turn all documents are re-assigned based on the new centroids. In general, if two data points are similar enough, they are considered as neighbors of each other. Every data point can have a set of neighbors in the data set for a certain similarity threshold [16]. Each text document can be viewed as a tuple with boolean attribute values, where each attribute corresponds to a unique term. An attribute value is true if the corresponding term exists in the document. Since a boolean attribute is a special case of the categorical attribute, we could treat documents as data with categorical attributes. With this assumption, the concepts of neighbors could provide valuable information about the documents in the clustering process. We believe that the intra-cluster similarity better be measured not only based on the distance between the documents and the centroid, but also based on their neighbors. The concept based correlation can then be used to enhance the evaluation of the closeness between documents because it takes the information of surrounding documents into consideration. This step iterates until converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids. A good candidate for an initial centroid should not only be close to a certain group of documents but also well separated from other centroids. By setting an appropriate similarity threshold ϴ, the number of neighbors of a document in the data set could be used to evaluate how many documents are close enough to the document.

5. With formation of Document Clusters, information retrieval is made more efficient by finding query and cluster similarity alone and indexing the retrieved documents.

The following experimental results demonstrate that the proposed method has better performance by increasing the number of relevant documents.

4. PERFORMANCE MEASURES

A. Experimental Setup

Here, we implemented this system using ASP.NET and C# 3.5 and SQL Server 2005 using sample IEEE journal papers of various domains as dataset. The documents are large dimensional data elements. At first, the dimension is reduced using the stop word elimination and stemming process. The system is tested with 300 sample IEEE journal papers of various domains such as data mining, networking, cryptography, multimedia, mobile computing, image processing etc., collected. For each article (document) in the corpus, the system used only its abstract for the evaluation. After preprocessing the system the term weighting is done and the term which has maximum frequency of occurrence is found and its synonyms are extracted from Conservapedia using web services. The extracted synonyms are stored in the database. Therefore, the target document corpus will be clustered in accordance with these concept represented ones and thus achieve the proceeding of document clustering at the conceptual level. This improves the accuracy of query based document retrieval and indexing from the document clusters.

Finally, it is important to measure the efficiency of the proposed method. The proposed method of the research adopted the most commonly used measures in the data mining, namely, precision and recall for the general assessment.

B. Cluster Quality Evaluation Measure

The external quality measure is the F measure, a measure that combines the precision and recall ideas from information retrieval. The higher f-measure is the higher accuracy of cluster, includes precision and recall.

Here, each cluster is treated as if it were the result of a query and each class is treated as if it were the desired set of documents for a query. Then, the recall and precision of that cluster for each given class is computed. More specifically, for cluster *j* and class *i*

Recall (i, j) = nij / ni

Precision (i, j) = nij / nj

Where,

nij is the number of members of class i in cluster j,

nj *is* the number of members of cluster j and

ni is the numbers of members of class *i*.
The F measure of cluster *j* and class *i* is then given by

F(i,j)=(2*Precision(i,j)*Recall(i,j))/((Precision(i,j)+ Recall (i, j)) …………………. (2)

Overall value for the F measure is computed by taking the weighted average of all values for the F measure as given by the following

$$Fc = \sum_i \frac{ni}{n} * \max\{F(i,j)\} \quad \text{………………. (3)}$$

n is the total number of documents.
Higher the value of F-measure better is the cluster quality.
Precision P is defined as the proportion of retrieved documents that are relevant [17].

$$P = \frac{Ra}{A} \quad \text{……………………... (4)}$$

Recall is defined as the proportion of relevant documents that are retrieved [18].

$$R = \frac{Ra}{R} \quad \text{……………………….. (5)}$$

A is the number of retrieved documents. R is the number of relevant documents. Ra is the number of retrieved relevant documents.

The experimental results obtained in the proposed system are as follows:

Table 1. Accuracy of the Proposed system.

Method of Clustering	Precision	Recall	F-measure
Without Synonym Extraction	0.7657	0.8025	0.7875
With Synonym Extraction	0.7788	0.8654	0.8231

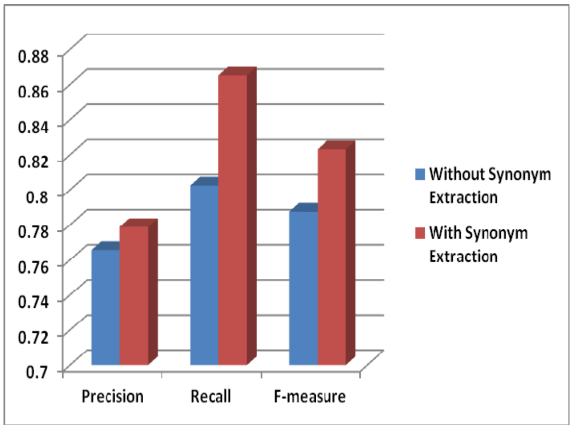

Figure 2. Experimental results based on accuracy.

5. CONCLUSION AND FUTURE WORK

The experimental results in Table 1 and Figure2 indicate that the proposed semantic synonym extraction method for document clustering increases the performance of information retrieval system in terms of accuracy by increasing the number of relevant documents retrieved than the traditional tf-idf model alone used for document clustering by K-means. In this work, few issues e.g. high dimensionality and accuracy are focused but there are still many issues that can be taken into consideration for further research. In future, this work can be extended in the direction of web documents clustering.

6. ACKNOWLEDGEMENT

We acknowledge the management of Sastra University for providing all the facilities for our research on this work. We would also like to thank Dr.B.Shanthi and Dr.N.Sairam, Professors, SOC, Sastra University, for their continuous encouragement.

REFERENCES:

[1] Pang-Ning Tan, M.S., Vipin Kumar, "Introduction to Data Mining". *Pearson International ed. 2006*, Pearson Education, Inc.

[2] M.A. Hearst, a.J.O.P. "Reexamining the cluster hypothesis," *In Proceeding of SIGIR '96*.

[3] Jardine, N.a.v.R., C.J., ".The Use of Hierarchical Clustering in Information Retrieval", *Information Storage and Retrieval.* Vol. 7, 1971.

[4] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Commun. *ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[5] Masoud Makrehchi, **"**Query-Relevant Document Representation for Text Clustering**",** *IEEE (ICDIM) 2010.*

[6] H. Sch¨utze and C. Silverstein, "Projections for efficient document clustering," *SIGIR Forum*, vol. 31, no. SI, pp. 74–81, 1997.

[7] G. A. Miller, "Wordnet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[8] Hotho, S. Staab, and G. Stumme, "Wordnet improves text document clustering," in *Proceedings of the SIGIR 2003 Semantic Web Workshop*. New York, NY, USA: ACM, 2003, pp. 541–544.

[9] Huang, D. Milne, E. Frank, and I. Witten, "Clustering Documents Using a Wikipedia-Based Concept Representation," in *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. Springer, 2009,pp. 628–636.

[10] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2000, pp. 208–215.

[11] Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 269–274, New York, NY, USA.

[12] Congnan Luo, Yanjun Li, Soon M.Chung , "Text document clustering based on neighbors", *Data & Knowledge Engineering, 2009.*

[13] Anton Leuski, "Evaluating Document Clustering for Interactive Information Retrieval", 2000.

[14] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques". *In KDD Workshop on Text Mining*, 2000.

[15] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering", In *Proceedings of the FifthACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1999.

[16] R. A. Jarvis and E. A. Patrick, "Clustering Using a Similarity Measure Based on Shared Nearest Neighbors," *IEEE Transactions on Computers*, Vol. C-22, No. 11, November, 1973.

[17] Jose A. Alonso-Jimenez, Joaquin Borrego-Diaz, Antonia M.Chavez Gonzalez, Francisco and J. MartinMateos, "Foundational challenges in Automated Semantic Web Data and Ontology Cleaning", *IEEE Intelligent Systems*, pp. 42-52, 2006.

[18] J. Han, and M. Kamber, "Data mining: Concepts and Techniques", *Morgan Kaufmann*, 2001.

AUTHOR PROFILES:

G.Bharathi received the Bachelor's degree in Computer Science and Engineering from Bharathidasan University, in 2002. Currently, she is doing her Master's degree in Computer Science and Engineering in Sastra University. Her interests are in Data mining, Text Mining and Information Retrieval.

Prof. D.Venkatesan has done his M.Sc in Applicable Mathematics and Computer Science and his M.Phil in Computer Science.
He is doing his PhD in Computer Science in Sastra University. Currently, he is an Assistant professor at Sastra University. His research interests include Soft Computing Techniques, DBMS and Data mining.