ISSN:0975-9646

Text Clustering for Information Retrieval System Using Supplementary Information

Chitra Kalyanasundaram[#1], Snehal Ahire[#2], Gaurav Jain[#3] , Swapnil Jain[#4]

Computer Department, K.K.W.I.E.E.R
Amruthdham, Nashik, India

Abstract- **Text clustering extends over wide range of applications from information retrieval system, pattern recognition, search engines to social networks, and other digital collections. Text data involved in such applications usually have ample of unused data associated with them. The paper focuses on handling this unused data, referred as supplementary information, to generate effective clusters. The supplementary information may include document provenance information, links in a document, index terms used within a document or any other data that is not generally used for clustering. In this paper, we perform document clustering using supplementary information along with the content for generating clusters with higher purity. We also identify the use of such supplementary information for clustering in applications involving other file types like audio, image, video, etc. The clustering performance may degrade if the supplementary information associated with pure content is noisy. Taking this into consideration, we use partitioning-based clustering algorithm and a probabilistic model. We present experimental result to justify the approach.**

Key Words- **Clustering, Information Retrieval System, partition-based clustering algorithm, Probabilistic model, Supplementary information, Similarity measure**

I. Introduction

Clustering is a technique in which data objects are divided into groups. The objects in the same group are similar in some context and those in different groups are dissimilar. Clustering applied to text domain is referred as text clustering. Text clustering finds applications in information retrieval systems, document organization and indexing, web, Customer segmentation, automated production of hierarchical taxonomies for browsing, etc. [1]

Clustering is an active research area. When used in the data mining domain clustering involves a lot of complications as it has to deal with a large datasets and a large number of heterogeneous attributes. The most popular application is the search engine's information retrieval system. This has necessitated the development of effective clustering algorithms which work to generate effective results making the best use of available data. Some of the techniques used for clustering so far are scatter-gather technique [5], which uses a combination of Agglomerative and partition-based clustering, Co-clustering method, Expectation Maximization (EM) method, Matrix-factorization [1]. These and other research work in this domain focuses on pure text clustering [6].

If real-time applications of clustering are considered, it is observed that the data to be clustered have a lot of extra attributes associated with them. Increasing use of social media and web has resulted in lot of additional information. Such data can be either meta-data associated with any document, Web-logs, links among documents, index terms used in the documents or any other data which are more informative and not usually used for clustering purpose. Such attributes will be referred as supplementary information as they will be used in supplement to the main content for generating pure clusters or increase recall factor in information retrieval system [2]. This supplementary information can be used in the clustering process to generate effective clustering. This results in rich browsing experience, retrieving relevant information, or well organized document depending on the application.

Additionally, text data can be associated with files of different types. In case of image file, the text data in this context may include camera details, location, or any other meta-information. Such information may be extracted and used for clustering as an alternative to image based clustering for some applications. Similarly, media files like audio or video also have Meta information which can be used for clustering.

Few examples which elaborate the concept of this paper are described next. Consider that papers published in a journal have to be clustered. Unlike existing approaches, instead of clustering them using the content of papers, supplementary information associated with the papers can be considered. Such information includes author name, key terms or index terms, links, references, etc. Using this information leads to effective clusters. The supplementary information can also be found with web documents which have meta-data associated with it. This meta-data is available in the form of web logs, provenance or other information about origin. While clustering text documents, there is a possibility that a document may contain a link to another. This kind of correlations among the documents cannot be found out by considering the pure text content. Such like can also be considered as supplementary information

The amount of data associated with dataset to be clustered is very large. All of them may contribute to generate effective clusters. Some part of it may add to noise. In order to leverage the use of supplementary information that are more discriminative, relevant and act as promising attributes for pure cluster generation, we use partition-based algorithm with probabilistic model to obtaining effective results.

II. Implementation

In order to cluster the data using its content along with supplementary attributes let us consider a collection of documents D as input to the clustering system. Along with the input dataset number of clusters to be generated, N, is also an input parameter. The implementation of the clustering process is as shown in Figure 1.

A. Full text extraction

From the given set of documents extract all the text content to obtain a list of all the distinct terms included in the documents. The terms obtained in this step are obtained from the text document or in other words they are pure content which are used in traditional clustering algorithms.

B. Preprocessing or cleaning

The list of distinct terms obtained in the previous phase include stopwords like 'the', 'a', 'an', etc. These stopwords play no role in the clustering process hence can be removed. Besides stemming is also being performed which will keep the root work and eliminate the modifications of those. The result of this phase will be processed further which will improve discriminatory power of attributes.

C. Extraction of Supplementary information

Supplementary information may be present in different forms depending on the application. In case of web it may be in the form of web logs, Document meta-data or provenance information or index terms in case of text documents, meta-information in case of media files, etc. This makes the system semi-supervised. Depending upon the form in which supplementary information is available the supplementary attributes are extracted.

D. Formation of Vector Space Model

Vector Space modelling is an algebraic model for representing models in terms of vectors of identifiers. It involves document indexing where content bearing terms are extracted from the document text, the indexed terms are weighted to enhance retrieval of document relevant to the user, and ranking the document with respect to the query according to a similarity measure.

Documents are represented as vectors:

$$d_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$

Dimension of a vector is equal to the number of distinct terms.

Term Frequency and Inverse document Frequency (TF-IDF) Computation: In order to represent documents in the vector form for further computations TF-IDF score is made use of. The tf-idf, ranks the importance of a term in its contextual document collection. Term frequency is calculated as a ratio of the number of occurrences of a word in its document to

The total number of words in its document which is a normalized frequency. The inverse document frequency is the log of the ratio of the number of documents in the corpus to the number of documents containing the given term. Inverting the document frequency by taking the logarithm assigns a higher weight to rarer terms. Multiplying together the TF and IDF metrics gives a new metric that place importance on terms frequent in the document and rare in the corpus.

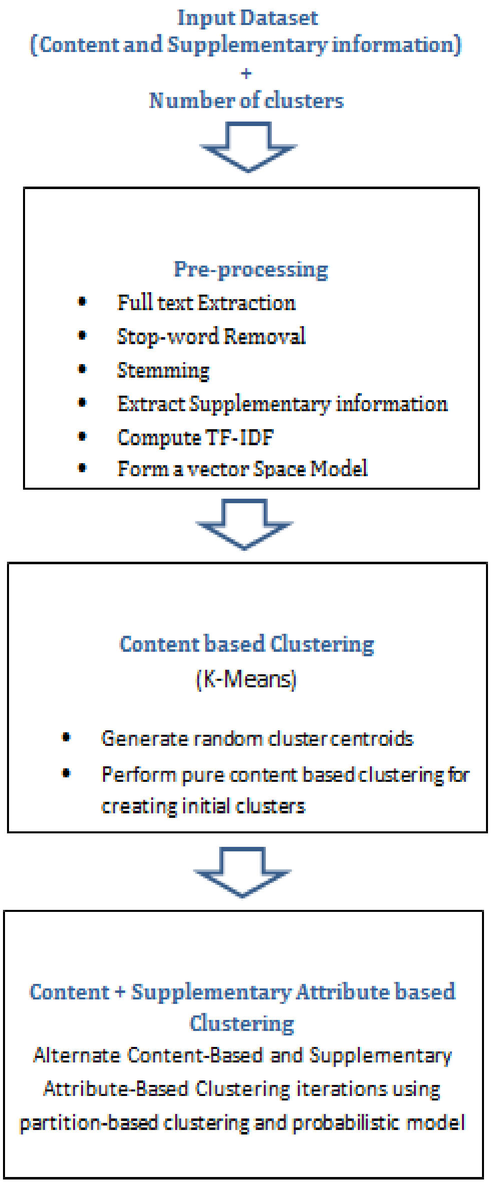
Input Dataset
(Content and Supplementary information)
+
Number of clusters
Pre-processing
• Full text Extraction
• Stop-word Removal
• Stemming
• Extract Supplementary information
• Compute TF-IDF
• Form a vector Space Model
Content based Clustering
(K-Means)
• Generate random cluster centroids
• Perform pure content based clustering for creating initial clusters
Content + Supplementary Attribute based Clustering
Alternate Content-Based and Supplementary Attribute-Based Clustering iterations using partition-based clustering and probabilistic model

Fig. 1 Clustering Implementation

E. Cluster Generation Process:

Along with the document vector for terms in a document a vector for auxiliary or supplementary attributes is also created. With the vector space model, the supplementary attributes and the number of clusters to be formed taken as input, an initial set of cluster centroids are created randomly. Using the cosine similarity function the documents in the corpus are assigned to the closest cluster centroid. This makes use of probabilistic model where the attribute probability is compared with cluster probability to assign it to closest cluster and assign cluster index to the documents. This results in the formation of initial clusters based on content of the documents. These initial clusters formed undergo multiple iterations during which clusters are refined [4].

The iterations use content and supplementary attributes alternatively so that each time the most recently formed cluster will be refined. Even single occurrence of a discriminatory supplementary attribute is very informative.

Sometimes information like link among two documents cannot be considered for clustering based on pure content. In such situation, the supplementary iteration occurring alternatively refines the most recent content based cluster appropriately. Considering most recent content based cluster ensures that the sole content is not being ignored while considering supplementary attributes.

As mentioned earlier the supplementary information may be noisy. For the clustering we consider only those attributes which take on the value 1 when considered in binary format. Those auxiliary attributes which are relevant take the value 1. The discriminatory power of supplementary attribute is computed using Gini Coefficient. Conditional probabilities of assignments of a document D to a cluster C is computed for those supplementary attributes which take the value 1 and are discriminatory. The one with highest value of probability is used for assignment. This will maximize effectiveness as well as suppress noise. Iterations terminate when the stopping criteria is met when cluster modifications is minimized to zero or a to a very small extent.

III. Results

The main motive behind our approach is to show that our model is better than natural clustering alternatives like simple k-means algorithm in terms of effectiveness of clusters generated. This has been achieved using pure content along with the supplementary information for cluster generation. In information retrieval contexts, precision and recall are defined in terms of a set of retrieved document and a set of relevant documents. We have used a combination of these two metrics to justify the results of our approach.

Recall is the ratio of the number of documents from class j in cluster i, to the number of documents in class j.

$$\text{Recall} = \frac{\text{Number of Documents from class j in cluster i}}{\text{Number of documents in class j}}$$

Precision is the ratio of the number of documents from class j in cluster i, to the number of documents in cluster i

$$\text{Precision} = \frac{\text{Number of Documents from class j in cluster i}}{\text{Number of documents in cluster i}}$$

F-Measure is a harmonic mean of precision and recall whose value ranges from 0 to 1.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

We have used a large dataset of IEEE transaction papers in order to demonstrate our results. The F-Measure has been computed for each cluster a with each class b and an average of most relevant cluster-class values has been computed and plotted as F-measure for different values of number of clusters. It can be observed that for the given dataset the F-Measure value for Text and Supplementary information based clustering is more than that of pure k-means clustering

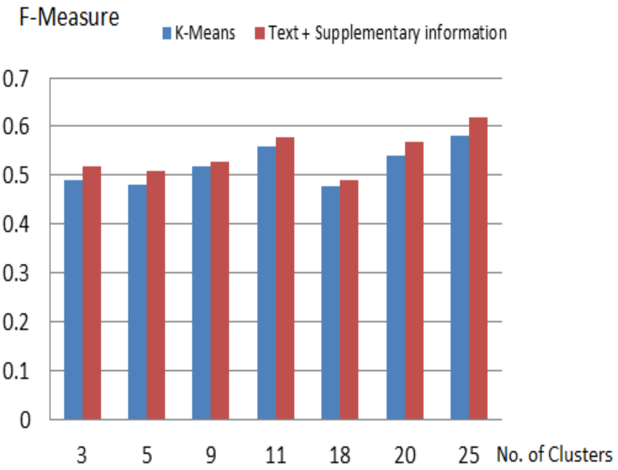

Fig. 1 Experimental Results

IV. Conclusions

In this paper, it has been shown how can the extra information associated with the data in different applications of text domain be used for the clustering. In information retrieval system, the retrieval speed is increased due to the use of clustering and use of supplementary information further increases the effectiveness of results obtained. By extracting supplementary data in text form available with other forms of data like image, audio, video, clustering can be done in a similar manner. Text clustering results have been derived in terms of precision and recall which shows the effectiveness of the approach when supplementary information is used.

Acknowledgment

Besides the four authors, other contributors to this work include Prof. S.S. Banait and Prof. Smita Patil from K.K.W.I.E.E.R.

References

[1]. C. C. Aggarwal and C.-X. Zhai, “*A survey of text classification algorithms,*” in Mining Text Data. New York, NY, USA: Springer, 2012.

[2]. IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 6, June 2014 “*On The Use Of Side Information For Mining Text Data* “Charu C. Aggarwal, Fellow, IEEE Yuchen Zhao, And Philip S. Yu, Fellow, IEEE

[3]. “*Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm*” Shi Na ; Liu Xumin ; Guan Yong Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on DOI: 10.1109/IITSI.2010.74 Publication Year: 2010 , Page(s): 63- 67

[4]. H. Schutze and C. Silverstein, “*Projections for efficient document clustering,*” in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 74–81.

[5] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “*Scatter/Gather: A cluster-based approach to browsing large document collections,*”in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.

[6]. M. Steinbach, G. Karypis, and V. Kumar, “*A comparison of document clustering techniques,*” in Proc. Text Mining Workshop KDD, 2000, pp. 109–110.

[7]. A. Kalton, K. Wagstaff, and J. Yoo, “*Generalized Clustering, Supervised Learning, and Data Assignment,*” Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, ACM Press, 2001