Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 2, Number 12; July-September, 2015 pp. 26-29

http://www.krishisanskriti.org/acsit.html

Facilitating Document Annotation for Efficient user Relevant Search

Sunil B. Ghane[1] and Sujata R. Kolhe[2]

[1]PG Student, Dept. of Computer Sci. &Engg. Datta Meghe COE Airoli-400708
[2]Dept. of Computer Sci. &Engg. Datta Meghe COE Airoli-400708
E-mail: [1]sunilghane83@gmail.com, [2]sujata.kolhe@dmce.ac.in

Abstract—*Document Annotation is the process of the addition of metadata information which is very important to fetch the information from the specific document. Due to the increasing sizes of data, it becomes cumbersome for anyone seeking for the information needs to provide assistance of automated approach to find what they are searching for. Clustering is a very important process to extract information from unstructured data, and enhance the process of grouping similar items together. Clustering also helps to discover hidden information and summarize a large amount of data into a small number of groups.*
We present an approach which maintain domain specific dictionary that helps to generate structured information by recognizing the documents which contain target specific information. This information is going to be useful for following process of querying database. To the input documents text pre-processing was initially done to extract the terms from the sentences and K-mean algorithm is used for clustering the documents. The goal of the system is to maximize the number of relevant documents in the ranked list as well as making sure that they are high up in the ranked list.

Keywords*: Clustering, Annotation, querying database, domain specific dictionary.*

1. INTRODUCTION

Day by day growth in technology makes it easy to access computing resources, which led to exponential amounts of information being generated over the past decade and is still growing. Information is stored, which is used for future reference, or use for local access and which can be used publicly. While search engines were developed to deal with this huge volume of documents, even they output a large number of documents for a user given query. Under these conditions it became very difficult for the user to find the documents, they actually needs, because most of the users are reluctant to make the cumbersome effort of going through each and every documents for information [1]. Thus systems that can automatically summarize one or more documents are becoming increasingly desirable.

Data that resides in fixed fields within a record or file is referred as a structured data. How we can present this data, annotate and classify these documents properly? Is a question of concern? Therefore Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents is important.

Due to the increasing sizes of these data collections, it becomes cumbersome for anyone seeking for the information needs to provide assistance of automated approach to find what users are searching for. The aim of the system is the retrieval of textual information, based on user's query.

Therefore search engines accept a query representing what the user is seeking for, and provide a ranked list of potential answers. The goal of the system is to maximize the number of relevant documents in the ranked list as well as making sure that they are high up in the ranked list. Retrieved results are tied enough with the accuracy of the query. Query is often inefficient due to a vague information need but could also be because of an ineffective representation with respect to the information collection.

Section 2 gives overview of Literature review, Section 3 describes Methodology Section 4 consist of Data analysis in which results are given as per flow. Section 5 gives Conclusion.

2. LITERATURE REVIEW

Collaborative adaptive data sharing platform (CADS), which is an "annotate-as-you-create" infrastructure that facilitates fielded data annotation[1], it encourage the annotation at the time of document creation. Supervised Learning of Semantic Classes for Image Annotation and Retrieval[2], time consuming and expensive. Social Tag Prediction[3] which unable to deals with heterogeneous users.

Pay-as-you-Go User Feedback for Dataspace Systems[4], used schema matching techniques but are costly and time consuming. A large amount of structured information is buried in unstructured text, Receiver Operating Characteristic (ROC) curves is use to estimate the extraction quality in a statistically robust way and show how to use ROC analysis to select the extraction parameters in a principled manner[5].

3. METHODOLOGY

In this paper, research has been done on live data which is been taken from google news. For clustering we have selected Technology, Entertainment, Business domains, and news are collected from related domains. Also domain specific dictionary is maintained which consist of domain specific keywords. At first news are collected and saved in text files.

Thereafter, Pre-processing is done to remove the stop words and finally stemming process is applied on different words i.e if the word in document is playing then the word after stemming would be play. Then K-mean clustering algorithm is used for clustering from word accuracy.

3.1 Domain specific Dictionary

First, we have maintain domain specific dictionary which contains keywords from each stream as shown in Fig. 1. In concept dictionary, 3 streams as Technology, Business and Entertainment are the domains.

Technology	Business	Entertainment
algorithm	account	action
application	agreement	actress
computer	bids	cinema
database	economy	composer
graphics	estimates	lyrics
hardware	stock	song
Internet	Trade	theatre

Fig. 1: Domain Specific Dictionary

3.2 Clustering

In this module , Document clustering uses domain specific dictionary to cluster the similar documents which has been given as input to system. Domain specific dictionary contains sets of words that describe the contents within the cluster and by using k-mean clustering algorithm clustering is done. K-mean algorithm works as follows. First, decide the number of clusters k.

1. Initialize the center of the clusters

2. Attribute the closest cluster to each data point

3. Set the position of each cluster to the mean of all data points belonging to that cluster

4. Repeat steps 2-3 until convergence

The architecture diagram represents the overall structure of the system. The data is collected from google news and clustering is done using k-mean algorithm. The collected news saved in a folder as text files, then given as input to the system.

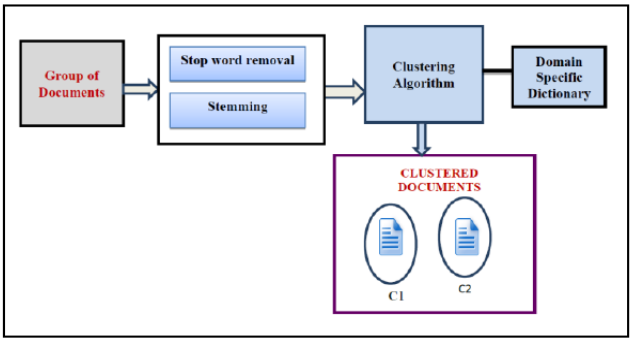

Fig. 2: System Architecture

4. DATA ANALYSIS

This section shows the results of implementation. Fig. 3 shows input selection process.

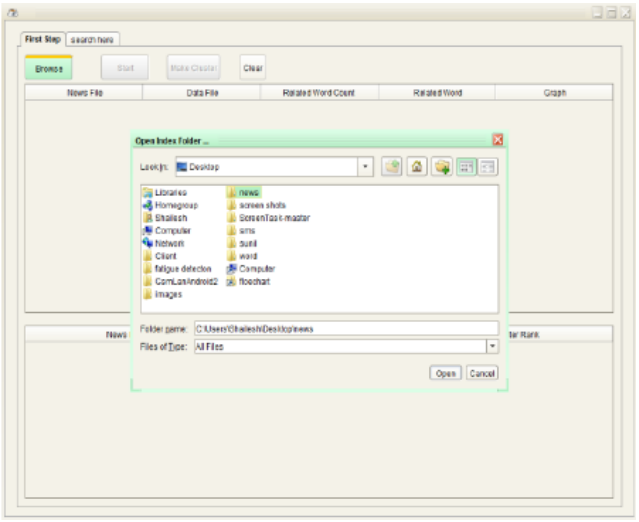

Fig. 3: Input selection

Fig. 4 shows the results, after applying stop words removal and stemming process on input text documents to find specific word accuracy to make cluster based on keywords.

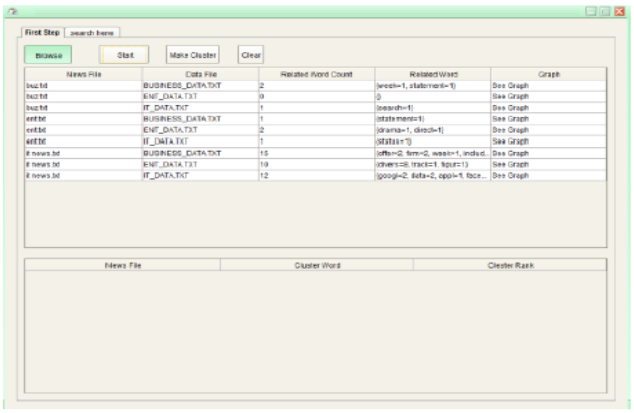

Fig. 4: Stemming

Fig. 5, shows result of word accuracy in graph for input text file. It displays the count of words from related domain.

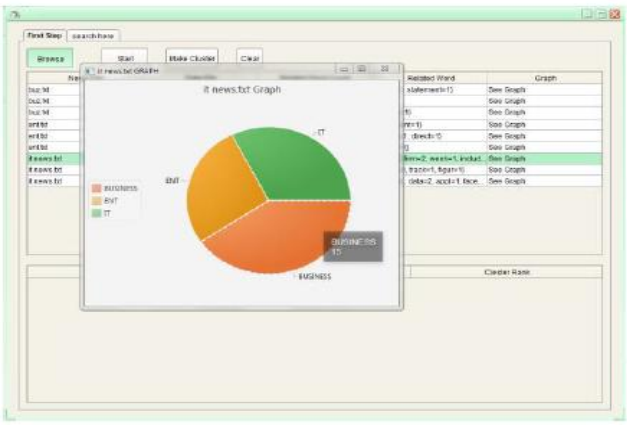

Fig. 5: Word Accuracy graph

Fig. 6 show the process of clustering using keywords.

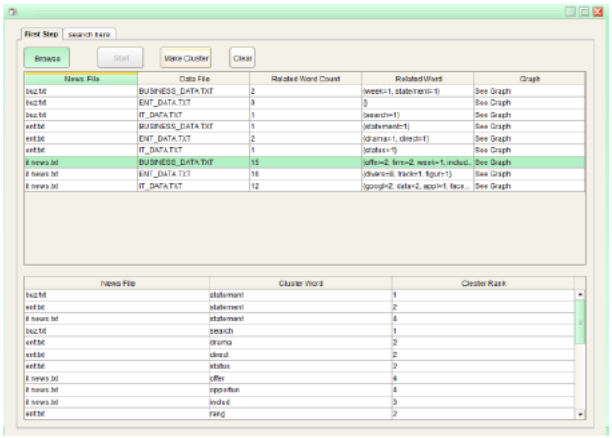

Fig. 6: Clustering

Fig. 7, shows searching process in which complete data is analysed to retrieve the results. It also shows precision and recall for the number of documents retrieved.

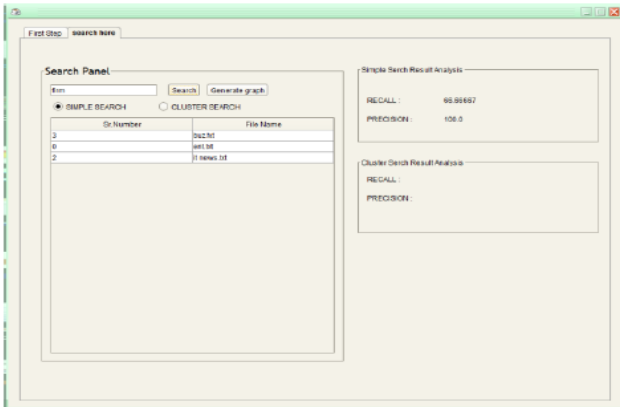

Fig. 7: Simple search

Fig. 8, shows searching process in which documents are searched, based on keywords within the cluster.

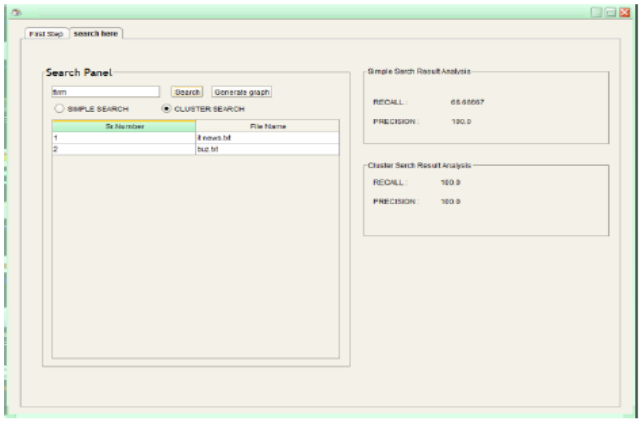

Fig. 8: Cluster Search

Fig. 9, shows graphical representation between simple search and cluster search with time difference between both searching techniques.

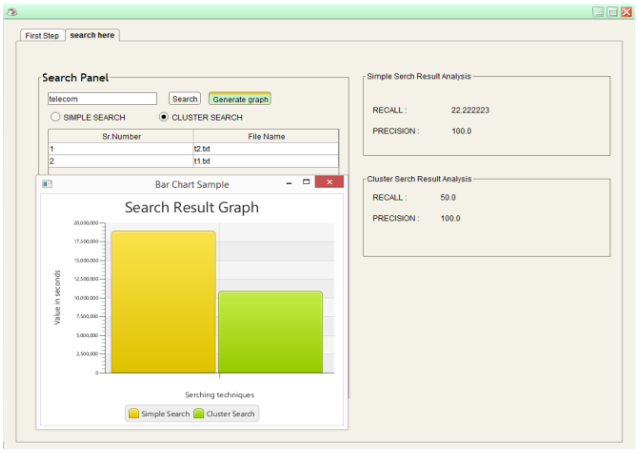

Fig. 9: Time Evaluation

5. CONCLUSION

We have used domain specific dictionary consists of keywords of related domains. Input documents text pre-processing was initially done to extract the terms from the sentences. Further stop words were removed and stemming was done and then K-mean algorithm is used for clustering the documents. The goal of the system is to maximize the number of relevant documents in the ranked list as well as making sure that they are high up in the ranked list.

REFERENCES

[1] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis (2014), "Facilitating Document Annotation Using Content and Querying Value", IEEE TRANSACTIONS, VOL. 26, NO. 2.

[2] G. Carneiro, A.B. Chan, P. Moreno, and N. Vasconcelos (2007), "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29-3, pp. 394-410.

[3] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina (2008), "Social Tag Prediction", SIGIR'08, July 20–24.

[4] S.R. Jeffery, M.J. Franklin, and A.Y. Halevy (2008), "Pay-as-You-Go User Feedback for Dataspace Systems," Proc. ACM SIGMOD Int'l Conf. Management Data.

[5] A. Jain and P.G. Ipeirotis (2009), "A Quality-Aware Optimizer for Information Extraction," ACM Trans. Database Systems, vol. 34, article 5.

[6] Y. Rui, T. S. Huang, and S. Mehrotra (1997), "Content-based image retrieval with relevance feedback in MARS," Proc. IEEE Int. Conf. Image Processing , pp. II815–818.

[7] Zheng Lu and Hongyuan Zha (2013), "A New Algorithm for Inferring User Search Goals with Feedback Sessions", vol. 25.

[8] D. Mclean, Y. Li, and Z.A. Bandar (2003), "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 871-882.

[9] A. Jain and P.G. Ipeirotis (2009), "A Quality-Aware Optimizer for Information Extraction," ACM Trans. Database Systems, vol. 34, article 5.

[10] Jiang Bian, Anlei Dong, Xiaofeng He, Srihari Reddy (2013), and Yi Chang, "User Action Interpretation for Online Content Optimization", vol. 25.

[11] M. Jayapandian and H.V. Jagadish (2008), "Automated Creation of a Forms-Based Database Query Interface," Proc.VLDB Endowment, vol. 1, pp. 695-709.

[12] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen (2004), "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248.