Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.778 – 782

RESEARCH ARTICLE

Document Cluster Mining on Text Documents

Twinkle Svadas[1], Jasmin Jha[2]

[1]Computer Engineering, L.J.I.E.T, India

[2]Computer Engineering, L.J.I.E.T, India

[1] svadas.twinkle23@gmail.com; [2] jhajasmine@gmail.com

Abstract— Along with the rapid and fast development of Internet, there is a tremendous increase in the use of online data and information. The exponential growth of data has led us to an information explosion era, where the data cannot be easily maintained. Also there is an increase in the use of electronic data and the information is stored in electronic format in the form of text documents such as news articles, books, digital library and so on. Clustering of the text documents has become an important technology over internet. Text Clustering is mainly described as grouping of the similar documents a large collection of unstructured documents. Text document clustering is the most widely used method for generalizing large amount of information. This paper proposes a system to categorize the text documents and form the clusters.

Keywords— Clustering, Document Clustering, Text Mining.

I. INTRODUCTION

With the wide use of internet, a large amount of textual documents are present over internet. Text data is present everywhere on the Web, in the form of enterprise information systems, digital documents and in personal files. As the size of text data is increasing at a surprising speed, the handling and analysis of text data becomes very important. Text mining is being developed as technology to handle the increasing volumes of the text data. Different text mining functionalities are text clustering, text classification, text categorization.

The use of communication technologies and the information content has increased extensively. It provides access to a large amount of data. Also databases are increasing in volumes day by day and they are of different types. In these data a decision making information is hidden which is needed to be analysed in order to obtain knowledge from the data. Data mining is process to generate patterns and derive relationships from unprocessed data using various analysis tools. Data mining refers to extracting knowledge from large chunks of data. Text Mining is the term used to describe either a single process or a collection of processes in which we obtain the previously unidentified information by automatically extracting information from different digital data sources. Text databases are also increasing due to rapidly growing online information in the form of electronic documents. The text databases contain the information in an unstructured manner. Applications of Text Mining are increasingly important with the growth in the production of personal and public information with the enhancement of web and social media. A large amount of information is present on the social media which is analysed with the help of text mining to generate meaningful patterns and latest trends.

Data Mining and its techniques are generally used to manage non numerical data. Clustering is a data mining technique that is typically used to create clusters from large amount of unstructured data sources which is the non numerical data. Clustering technique has been used in many of the data mining problems such as to build relations from a complex dataset, to find associations between the objects and to make generalizations. Clustering applications have been applied to a large variety of areas ranging from engineering, life and medical sciences, social science, computer science, economics and so on. The applications of clustering has also increased in fields like information retrieval, text mining, web applications, spatial database analysis and analysis of DNA in the field of biology. Traditional clustering methods were applied to the numeric data and were developed in the statistical context. Now the non numerical that is textual data analysis has been introduced. For the treatment of the textual data it is worth to compute the semantics related between the terms and concepts present in the data. The semantics can be obtained using some knowledge base or sources like dictionaries, web, textual corpus or ontology. The use of knowledge base provides better management of the textual data at the semantic level. As a result the use of knowledge base with the clustering methods may improve the results when we are dealing with the textual data.

In the past decade, the use of World Wide Web has tremendously increased. As a result, the web is growing day by day at a very fast pace. Information search using web has become very common nowadays. A large amount of information is present on the web in the form of different data sources like e-library, digital publishing. Hence organizing the information on web helps the user to retrieve meaningful information. Data Mining on web thus becomes useful for extracting useful knowledge from web. Data Mining is a technique which retrieves or extracts meaningful knowledge from large amount of data. Semantic Web has also become an important research area. With the use of digital technologies the web has become a vast source of information. Looking for the accurate, relevant and precise information from the web and extracting it from the web has become a tedious task. Many techniques are present for information extraction from web and text mining is one such technique.

Text Clustering, a data mining technique can be used along with ontology to group the similar digital data which enhances the clustering process. With the growth of internet, conventional newspapers have been developing their presence on web. A large amount of online news data is present on web. Text Mining can be used to extract the similar news articles from the web. Social Media is nowadays one of the most updated forms of writing news articles. The increasing volume and availability of large amounts of online data in social media environments provides new opportunities for researchers to monitor social and economic behaviour. Data Mining can be used to extract the similar news articles from the web and cluster them on the basis of concept weight and similarity measures. A lot of work has been done on text clustering, but combining text clustering and web ontology is an emerging area of research. The domain ontology can act as a knowledge base and can be used to calculate the semantics between the concepts or terms and can be used with clustering.

In this paper a text clustering system is presented. This system consists of the different components like preprocessing, clustering and ontology.

II. RELATED WORKS

Jian Ma, Wei Xu [1], proposes an Ontology based Text Mining (OTMM) method to cluster research proposals in a research funding agency. An ontology in the domain of research is created to categorize different research areas. The proposals are then classified into different disciplines and a text clustering algorithm, Self-organized Mapping, is applied to cluster the research proposals on the basis of similarity. After the grouping, the proposals are assigned to the reviewers. Hence this approach reduces the time of grouping the research proposals and assigning to the intended reviewers and promotes efficiency in proposal grouping process.

S. C. Punitha, M. Punithavalli[3], studied two approaches for text clustering and compared them. First method is based on pattern recognition with semantic driven methods for clustering text documents. Second method is an ontology based text clustering approach. Both algorithms are analysed in terms of efficiency and speed of clustering. Experiments proved that both techniques were efficient in clustering process, but the performance of ontology based approach was better in terms clustering quality, but a relatively slow speed because of more computations.

E. Alan Calvillo, Alejandro Padilla[2], proposes a method to cluster research papers by using text clustering. The K-Means algorithm is used to implement the semi-supervised learning clusters to identify and approximate search using as per defined pattern. The limitation of this paper is that it applies semi automatic learning from a knowledge base. An automatic learning process can be applied that can enhance the search from the manipulated texts. Such techniques can be applied to the database knowledge with the help of filter, wrapper and even ontology.

Qiujun LAN[4], proposes an approach to for extraction of news content using similarity measure based on edit distance to separate the news content from noisy information. This paper describes about the accurate extraction of news content from web pages. A backward and forward similarity measure is used based on edit distance method. The algorithms used with this method are less complex with high accuracy and efficiency rate. It is appropriate method to extract news content from noisy data in news web mining.

TABLE I
COMPARISION OF RELATED WORKS

Introduction	Working	Conclusion	Problem & Future Work
News keyword extraction for topic tracing [5].	Keyword extraction using TF-IDF & its variants.	Extract main features from document set & filter keywords with cross domain comparison.	Feature extraction from subjective documents for user review.
Pattern and cluster mining on text data [6].	Document cluster using k-means & hierarchical agglomerative clustering.	Association among words using given algorithm	Applying probabilistic model for selection of feature vectors.
Classification of news document [8].	Text categorization algorithm based on Bracewell method.	Good Accuracy is obtained in offline and online mode.	To integrate with clustering for better categorization & more efficient computation.
Classification of news articles TF-IDF approach [7].	Classification using TF-IDF algorithm.	TF-IDF could classify articles with good accuracy. The accuracy of different categories is diverse.	Improper categorization and takes lot of time.
An Ontology-Based Text-Mining Method to Cluster proposals for Research Project Selection [1].	Clustering research proposals using SOM Algorithm	Group Research Papers according to categories and assign to reviewers	To Cluster external reviewers and assign research proposals systematically.
Searching Research Papers Using Clustering and Text Mining [2].	Clustering research papers using KMeans Algorithm	To optimize information and fast Searching	To implement Automatic learning

III. ARCHITECTURE OF THE PROPOSED SYSTEM

The proposed system is designed cluster the text documents. It consists of the following components like preprocessing and ontology. Fig. 1 shows the architecture of the proposed system. The functionality of the components is as described below.

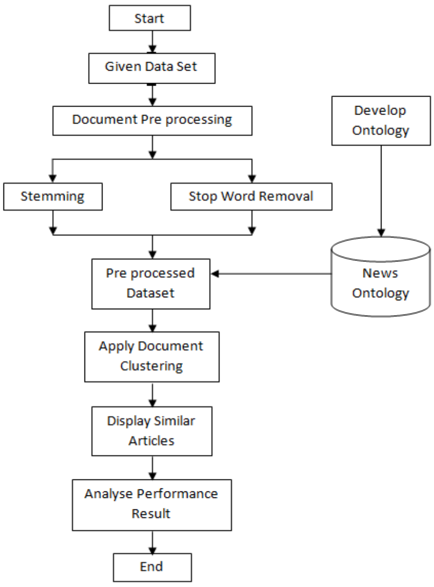


Fig. 1 Architecture of the proposed system

Preprocessing: It involves all processes, methods that are required to prepare data for text mining. It converts data from original form to machine readable format before applying feature extraction methods to generate new collection of documents represented by the concepts. Techniques like stop word removal, stemming and tokenization are involved in preprocessing.

- **Stop word removal:** Very often a common word, which would appear to be less significant in selecting document that would match a user's need, is completely expelled from the vocabulary. Such words are called stop words and the technique is called stop words removal technique. This technique increases the effectiveness and efficiency. Example of the stopwords are a, is, the, when, etc.

- **Tokenization:** Tokenization is the process of breaking up given character sequences into meaningful words, symbols, or crunches of data while maintaining its security and integrity which can be further used for processing.

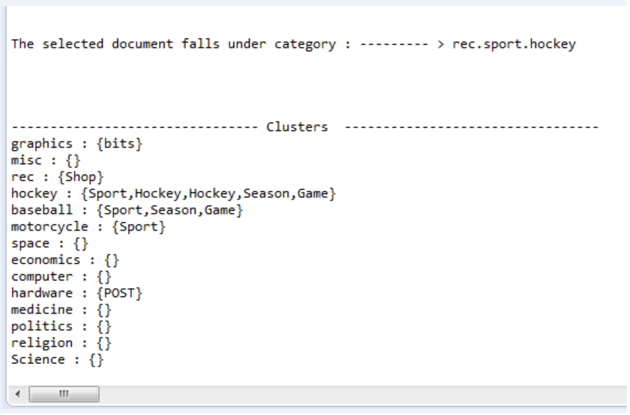
Clustering: Text Clustering is the application of the data mining functionality, of cluster analysis, to the text documents. Document or text clustering is an important technique to organize documents. Text Clustering helps to cluster similar kinds of digital documents. This method is used on web to cluster digital data to enhance the search and to retrieve meaningful lists of the data. Various clustering algorithms are present to cluster similar objects into one cluster and dissimilar objects into different cluster.

Ontology: Ontology can be considered as a repository of knowledge in which concepts and terms are defined and also the relationships between these terms and concepts are given. It is a set of concepts and relationships that describe a domain of interests and represents an overview of the domain. Ontology makes the knowledge that is implicit for humans, explicit for computers [1]. Hence ontology automates information processing and can improve text mining for a particular domain.

IV. EXPERIMENTS

A large collection of text documents are considered as unstructured data. It is very difficult to group the text documents. A dataset is used for the clustering of documents. For this purpose 20 News Group Dataset is used. The dataset consists of collection of large number of documents. It consists of a collection of 20000 documents partitioned into 20 different categories. Data from this document collection is taken as input. The documents from this collection are chosen at random for the experiments.

Pre processing techniques are applied to the dataset in order to obtain the pre processed dataset. Tokenization and stop word removal techniques are applied. Clustering is then applied to the collection of text documents. Fig. 2 shows the clusters of the documents from different categories.



```
The selected document falls under category : --------- > rec.sport.hockey



------------------------------- Clusters  --------------------------------
graphics : {bits}
misc : {}
rec : {Shop}
hockey : {Sport,Hockey,Hockey,Season,Game}
baseball : {Sport,Season,Game}
motorcycle : {Sport}
space : {}
economics : {}
computer : {}
hardware : {POST}
medicine : {}
politics : {}
religion : {}
Science : {}
```

Fig. 2 Cluster Formation

The proposed method used the most common approaches for the general assessment like precision, recall and accuracy. This is further illustrated in the fig. 3.

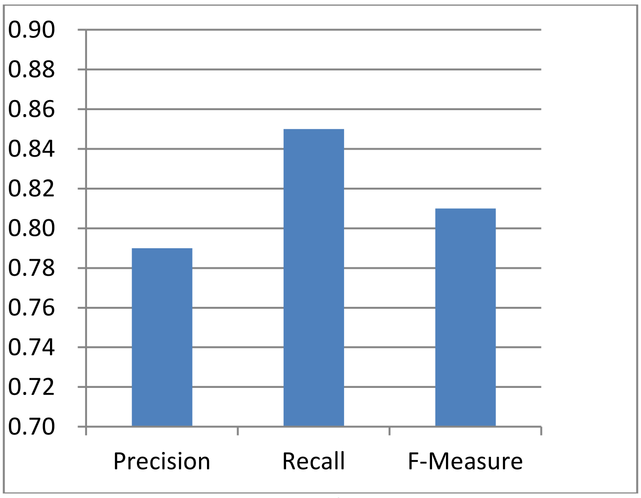


Fig. 3 Results of experiment

V. CONCLUSION AND FUTURE WORK

Various Text Clustering Algorithms can be used to categorize news articles. The document collection is obtained and pre processing techniques are applied to the document collection in order to remove stop words and to do tokenization. This would remove unnecessary words from the document collection and would provide a pre processed dataset. A document clustering algorithm is used for clustering and categorizing the news articles on the basis of topic. Mining similar news articles from web and applying ontology based text clustering algorithm provide clusters of similar news articles. The results of clustering are improved by using the ontology based clustering algorithms rather than simple clustering algorithms. As future work such system can be used to work for big data.

REFERENCES

[1] Ma, J., Xu, W., Sun, Y. H., Turban, E., Wang, S., & Liu, O. "An ontology-based text-mining method to cluster proposals for research project selection". Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 42(3), 2012, ISSN 1083-4427, pp.784-790.

[2] Calvillo, E. A., Padilla, A., Munoz, J., Ponce, J., & Fernandez, J. T. "Searching research papers using clustering and text mining". In International Conference on Electronics, Communications and Computing (CONIELECOMP), 2013, IEEE, ISBN 978-1-4673-6156-9, pp. 78-81.

[3] Punitha, S. C., and M. Punithavalli. "Performance Evaluation of Semantic Based and Ontology Based Text Document Clustering Techniques." International Conference on Communication Technology and System Design, Procedia Engineering 30, Science Direct, Elsevier 2012, DOI 10.1016/j.proeng.2012.01.839, pp. 100-106.

[4] Qiujun, L. 2010. "Extraction of News Content for Text Mining Based on Edit Distance", Journal of Computational Information Systems, 2010, pp.3761-3777.

[5] Lee, S., & Kim, H. J. "News keyword extraction for topic tracking". Fourth International Conference on Networked Computing and Advanced Information Management, 2008. NCM'08, IEEE, ISBN 978-0-7695-3322-3,Vol. 2, pp. 554-559.

[6] Agnihotri, D., Verma, K., & Tripathi, P. "Pattern and Cluster Mining on Text Data". Fourth International Conference on Communication Systems and Network Technologies (CSNT), 2014, IEEE, ISBN 978-1-4799-3069-2, pp. 428-432.

[7] Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W."Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach". 6th International Conference on Information Technology and Electrical Engineering (ICITEE), 2014, IEEE, ISBN 978-1-4799-5302-8, pp. 1-4.

[8] Rachmania, A., Jaafar, J., & Zamin, N." Likelihood calculation classification for Indonesian language news documents". International Conference on Information Technology and Electrical Engineering (ICITEE), 2013, IEEE, ISBN 978-1-4799-0423-5, pp. 149-154.