Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 100 – 106

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

Performance Evaluation of Semantic Based and Ontology Based Text Document Clustering Techniques

S. C. Punitha[a], M. Punithavalli[b], a*

[a] *Department of Computer Science, P.S.G.R Krishnammal College For Women, Coimbatore,Tamilnadu,India.*
[b] *Computer Science Department, Sri Ramakrishna college of Arts and Science for Women, Coimbatore, Tamilnadu, India.*

Abstract

The amount of digital information is created and used is steadily growing along with the development of sophisticated hardware and software. This has increased the need for powerful algorithms that can interpret and extract interesting knowledge from these data. Data mining is a technique that has been successfully exploited for this purpose. Text mining, a category of data mining, considers only digital documents or text. Text Clustering is the process of grouping text or documents such that the document in the same cluster are similar and are dissimilar from the one in other clusters. This paper studies the working of two sophisticated algorithms. The first work is a hybrid method that combines pattern recognition process with semantic driven methods for clustering documents, while the second uses an ontology-based approach to cluster documents. Through experiments, the performance of both the selected algorithms is analyzed in terms of clustering efficiency and speed of clustering.

Keywords :Dataming;Document clustering; HSTC; Feature Selection ;TCFSmethod.

1. Introduction

The information and communication industry has envisaged a dramatic increase in the amount of information or data being stored in electronic format. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making [1]. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations [2].

* S. C. Punitha. Tel.: +91-9362699988.
E-mail address: saipunith@yahoo.co.in.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.01.839

Data mining is a multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing and data visualization [3]. Even though, many researchers have probed into the field of data mining, it still has to go a long way for perfection. As the demand of customers grows the need for understanding the data and predict the future becomes crucial. In general, data mining basically performs three operations. They are (i) explore the data (ii) find patterns and (iii) perform prediction. To perform these steps, a number of data mining methods including data characterization, data discrimination, association analysis, classification, prediction and clustering are available.

Out of these techniques, a clustering-based approach to discover knowledge from text documents is taken as the topic of discussion in this paper. Document clustering or text clustering is a subset of the larger field of data clustering and text mining. The field borrows concepts from the fields of Information Retrieval (IR), Natural Language Processing (NLP) and Machine Learning (ML) field. The process of document clustering is to automatically group a document into a list of meaningful categories, in such a way that the documents in a category are similar to each other and dissimilar to documents in other categories [4].

Document clustering is the task of automatically organizing text document into meaning full cluster or group, such that the document in the same cluster are similar, and are dissimilar from the one in other clusters [5]. It is one of the most important tasks in text mining. There are several number of technique launched for clustering documents since there is rapid growth in the field of internet and computational technologies, the field of text mining have a abrupt growth, so that simple document clustering to more demanding task such as production of granular taxonomies, sentiment analysis, and document summarization for the scope of devolving higher quality information from text. They involve in multiple interrelated types of objects. Co-cluster means document similarity and word similarity are defined in a reinforcing manner.

Different approaches to solve the problem of document / text clustering have been proposed Decherchi *et al*. (2009) proposed a hybrid scheme that combined pattern recognition grouping algorithm with semantic driven method to arrange unstructured documents into content-based homogeneous groups[6]. This model is referred as HSTC (Hybrid Scheme for Text Clustering) in this paper. They used a semantic-based metric measure distance to calculate the similarity ratio between documents by performing a content and behavioral bases analysis. This had the advantage of taking into account the lexical and structural properties along with the style characteristics of the processed documents. They used a Radial Basis Function (RBF) for clustering. Another work that used semantic characteristics for text document was proposed by Raja and Narayanan (2010) and Thangamani and Thangaraj (2010). The model used a new Text Clustering with Feature Selection (TCFS) method to improve text document clustering. The system was designed to identify the semantic relations using ontology, which represents the term and concept relationship [7]. From these relationships, a concept weight is calculated and used during clustering. Both the systems offer efficient methods that enhance the document clustering process. This paper compares the performance of both these systems. The Reuters 21578 news document dataset is used to test their performance.

The remaining of the paper is organized as below. Section 2 describes the general document clustering process. Section 3 explains the methods and techniques used in HSTC scheme, while Section 4 explains the TCFS method. Section 5 presents the results obtained while testing both the systems with Reuters dataset. Section 6 concludes the work with future research directions.

2. The Document Clustering Process

Clustering algorithms in text mining are designed to discover groups in the set of documents such that the ones within a group are more similar to one another than to those belonging to other groups. The problem of document clustering can be described as below.

> *"Let $D = \{d_1, d_2, ..., d_n\}$ be a set of documents with $C = \{c_1, c_2, ..., c_m\}$ set of categories and $T = \{t_1, t_2, ..., t_n\}$ terms. Given an similarity or distance metric along with a partitioning criteria, cluster the documents into groups with similar features".*

Most of the clustering techniques aim to solve the above problem in a time efficient manner with maximum accuracy and belong to either a flat architecture or hierarchical architecture. The techniques that cluster documents without the need of document structure are termed as 'Flat' clustering. There are two types of flat clustering technique, one that requires the number of clusters, K, in advance (Manning *et al*., 2008) and another which can determine this number automatically (Ridella *et al*., 1998). Irrespective of the number, a membership function, Ω, that maps a document, d_i, to a cluster (1 - K) is used to minimize the partitioning cost with respect to the similarities among the documents. Another technique, called hierarchical clustering, groups documents in a structural, multilevel fashion and does not require the predefined value K, as it utilizes a series of partitioning tasks that finally results with a hierarchy of groups. Irrespective of the technique, when applied to text clustering, three issues should to be considered. They are, (i) Dimensionality (ii) Clustering process and (iii) Clustering algorithm.

In text clustering, the documents are represented using vector space models which treat a document as a bag of words. The bag of words approach increases the dimensionality of the feature space, which imposes a big challenge to the performance of clustering algorithms. Most of the clustering algorithms aim to reduce this high dimensionality while maintaining the document's semantic structure. Methods like spectral clustering [11], latent semantic index [12], locality preserving index [13], and non-negative matrix factorization [14] have been frequently used. All these methods have their advantages and disadvantages and have to be tuned up according to the application. Another fact that is worth noting is that not all terms or features collected are important during document clustering. There may be redundant or irrelevant data, which may influence the clustering process in a negative manner. Thus selection of quality features for clustering is important in terms of data understanding, clustering efficiency and dimensionality reduction [15].

Clustering process is the process of calculating a similarity measure that denotes the content similarity between two term vectors of two documents. The result is often used by the partitioning algorithm and is critical for obtaining quality clusters. The frequently used similarity measure is the 'cosine similarity' (Equation 1) which represents similarity as the correlation between the document vectors representing them.

$$\text{Cosine}(d_i, d_j) = \frac{d_i \bullet d_j}{\| d_i \| \| d_j \|} \tag{1}$$

where $\bullet$ represents vector dot product and $\| d_i \|$ is the length of vector d_i. The cosine value is 1 when two documents are identical and 0 otherwise. A larger cosine value indicates that these two documents share more terms and are more similar. With the result of the similarity measure, the next step is the actual clustering process. A variety of clustering algorithms are available which includes k-means, EM (Expectation Maximization) algorithm, Self Organizing Maps (SOM), fuzzy clustering.

A document clustering algorithm performs knowledge extraction or information extraction through the use of a series of sequential steps. Given a document dataset, the first steps perform a pre-processing to reduce the sequence of terms that are used to represent a document D by eliminating irrelevant data. The result produces a set of terms from which index term vector space can be generated that can be directly used by a machine learning algorithm. This process is called feature extraction or information extraction. The most frequently used model is the vector space model, which can be descried as follows. Given a collection of documents D, the vector space model represents each document 'd_i' as a vector of real-valued weight terms $v = \{w_j, j=1, ..., n_T\}$. Here w_i is a non-negative weight denoting the relevance of the term 'j' within a document containing 'n' terms.

3. Hybrid Scheme for Text Clustering (HSTC) Model

This model takes advantage of content-based processing for efficient clustering. The algorithm performs clustering on a dataset D containing 'n' documents represented as $D = \{D_j; j = 1 .. n_D\}$ having a collection of terms $T = \{t_i; i = 1 ..n_T\}$ obtained after performing pre-processing. The preprocessing performs stop-word removal and stemming to removal repeated and irrelevant terms. A content-based distance measure is used as similarity measure. This measure combines the distribution-based measure with the behavioural characteristics of the document features. The inclusion of behavioral characteristics includes document structure and style information into

similarity evaluation, so as to improve the overall clustering performance. While calculating the document distance measure, a document 'D' is represented using two vectors, V* and V**. V*(D) represents the content description of D and is a set of terms where each term 't' is associated with its normalized frequency 'tf'. Thus, the k^{th} element of vector $V^*(D_i)$ can be calculated using Equation (2).

$$V^* = tf_{k,i} \Big/ \sum_{l=1}^{n_r} tf_{l,i} \tag{2}$$

where $tf_{k,i}$ is the frequency of the k^{th} term in document D_i. Thus V* represents a document as a vector using term frequencies to set weights associated to each element. The distance between a pair of documents (D_i, D_j) is calculated using Equation (3) and is represented as $\Delta^{(f)}$.

$$\Delta^{(f)}(D_i, D_j) = \left[\sum_{k=1}^{n_r} | V^*_{k,u} - V^*_{k,v} |^p \right]^{1/p} \tag{3}$$

In the HSTC model, p = 1 and therefore actually implements Manhattan distance metric.

The second vector V** takes into consideration the structural properties of a document and is represented as a set of probability distributions associated with the term vector. Here, each term $t \in T$ occurring in a document D is associated with a distribution function that gives the spatial probability density function (pdf) of 't' in D. Such a distribution, $p_{t,u}(s)$, is generated under the hypothesis that, when detecting the k^{th} occurrence of a term 't' at the normalized position $s_k \in [0,1]$ in the text, the spatial pdf of the term can be approximated by a Gaussian distribution centered around s_k. In other words, if the term t_j is found at position s_k within a document, a second document with similar structure is expected to include the same term at the same position or in a neighbourhood thereof, with a probability defined by a Gaussian pdf. To derive a formal expression of the pdf, assume that the i^{th} document, D_i, holds n_o occurrences of terms after simplifications. If a term occurs more than once, each occurrence is counted individually when computing n_o, which can be viewed as a measure of the length of the document. The spatial pdf is defined using Equation (4).

$$p_{t,u}(s) = \frac{1}{A} \sum_{k=1}^{n_o} G(s_k, \lambda) \tag{4}$$

where A and λ are normalization terms, G is the Gaussian pdf given by Equation (5)

$$G(s_k, \lambda) = \frac{1}{\sqrt{2\pi}\lambda} \exp\left[-\frac{(s - s_k)^2}{\lambda^2} \right] \tag{5}$$

From this the second term vector V** is calculated by considering a discrete approximation of Equation (4). Here, the document D is segmented evenly into S sections, from which S-dimensional vectors are generated for each term $t \in T$. Each element estimates the probability of a term't' occurring in the corresponding section of the document. Thus, v**(D) is represented as an array of n_T vectors having dimension S. The distance between the probability vectors thus created (V**) is calculated by using Euclidean metric (Equation 6) and is represented as $\Delta^{(b)}$ for two documents D_i and D_j.

$$\Delta^{(b)}(D_i, D_j) = \sum_{k=1}^{n_r} \Delta^{(b)}_{t_k}(D_i, D_j) = \sum_{k=1}^{n_r} \sum_{s=1}^{S} \left\| v^n_{(k)s,i} - v^n_{(k)s,j} \right\| \tag{6}$$

From the calculated $\Delta^{(f)}$ and $\Delta^{(b)}$, the final distance is calculated using Equation (7).

$$\Delta(D_i, D_j) = \alpha\, \Delta^{(f)}(D_i, D_j) + (1-\alpha)\, \Delta^{(b)}(D_i, D_j) \tag{7}$$

here $\alpha \in [0, 1]$ is the mixing coefficient weight.

For the last stage, that is the actual clustering, a kernel-based k-means partitioning algorithm [16] is used for grouping similar documents in a top-down hierarchical process. In particular, a k-means clustering adopting rbf-kernel (radial basis function kernel) is used. A detailed description is given in Decherchi *et al*., 2009.

4. Text Clustering with Feature Selection (TCFS) Method

The methodology used by TCFS method is similar to that of HSTC method, but it differs in three ways. The first is in the preprocessing stage, second is the clustering process and the third is in the clustering algorithm used. Both use the same similarity measure, cosine distance. In the preprocessing stage, apart from stop word elimination and stemming, a weight estimation function, that calculates the term weight and semantic weight, are included. Term weight is estimated using TF/IDF values that utilize information about term and number of times (n) it appears in the document. Using the term weight value a term cube is constructed. A term cube is a 3-D model representing the document, term and n relationship. The semantic weight is calculated by concept extraction, concept or semantic weight calculation and construction of semantic cube. The concept extraction module is designed to identify concept in each document. This process is done with the help of the ontology collection. The terms are matched with concepts, synonyms, meronyms and hypernyms in the ontology. The concept weight is estimated with the concept and its element count. The semantic cube is constructed with concepts, semantic weight and document. In cluster processing which groups the documents, two techniques, namely, term clustering and semantic clustering technique are used. Term clustering groups documents according to the term weight, while semantic clustering groups documents according to the semantic weight. For clustering, a classical k-means algorithm is used.

5. Experiment Results

This section explains the results obtained while analyzing the performance of the two clustering models considered in this paper.

5.1. Reuters 21578

Reuters-21578 is the most widely examined text corpora from text mining. In has a collection of 21578 real-world news stories and news-agency headlines in the English language. Each of these articles is assigned to one of the 135 categories available. It is a freely available collection and is distributed as 22 files, each consisting of up to 1000 documents. The meta-data available for each document includes Date (of creation), Topics (a list of category labels) and Author. The text part of each document consists of a Title (the headline of the story) and Body (the content) section. A typical document is shown in Figure 1. More information about Reuters-21578 can be found at http://www.daviddlewis.com/resources/testcollections/reuters21578/ readme.txt.

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">
<DATE> 2-MAR-1987 16:51:43.42</DATE>
<TOPICS><D>livestock</D><D>hog</D></TOPICS>
<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>
<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.
Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudo rabies virus) control and eradication program, the NPPC said.
A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter
&#3;</BODY></TEXT></REUTERS>

Figure 1: Typical Reuters Document

5.2. Performance Metrics

To evaluate the performance of the two models selected in this study, two performance metrics, namely, F-measure and CPU execution time are considered. The F-measure is calculated from two measures, precision and recall, which are derived from four values, namely, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) during analysis of performance (Figure 2).

	Same category	Different categories
Same cluster	TP	FP
Different cluster	FN	TN

Figure 2: Confusion Matrix

The equation used to calculate precision (p) and recall (r) are given in Equations 8 and 9.

$$P(i,j) = \frac{N_{ij}}{N_j} \tag{8}$$

$$R(i,j) = \frac{N_{ij}}{N_i} \tag{9}$$

where N_{ij} is the number of objects of class 'i' in cluster 'j'. N_j is the number of objects of cluster 'j', N_i is the number of objects of class 'I'. The F-measure is calculated using Equation 10.

$$F(i,j) = \frac{2P(i,j)R(i,j)}{P(i,j)+R(i,j)} \tag{10}$$

The global F-measure for the whole clustering result is obtained using Equation 11.

$$F = \sum_i \frac{N_i}{N} \max_j (F(i,j)) \tag{11}$$

where N is the total number of documents in the data set. It is always desired to obtain a large F-measure, which indicates better clustering performance.

The CPU execution time is the execution time taken to complete the clustering process and can be used as a measure to measure efficiency and scalability of the algorithm while using a large dataset. The experiments were conducted using a Pentium IV machine with 2GB RAM.

Table 1 shows the results obtained with respect to F-measure and the execution time taken by the two selected algorithms

Table 1: Average F-Measure and Execution Time (Minutes)

Algorithm used	**F-measure**	**Execution Time**
HSTC	0.68	78.43
TCFS	0.71	79.66

From Table 1, it can be seen that the inclusion of ontology with clustering improves the performance of clustering by 4.2 per cent. While taking execution time into consideration, the TCFS algorithm is slightly slower than HSTC algorithm by 1.23 seconds. This might be due to the extra computations that need to be performed during term and semantic weight calculations.

6. Conclusion

As the volume of information continues to increase, there is growing interest in helping people better find, filter and manage these resources. Text clustering, which is the process of grouping documents having similar properties based on semantic and statistical content, is an important component in many information organization and management tasks. In the present research work two novel approaches to document clustering was considered and their methods and performance were analyzed. The first approach, HSTC, uses a hybrid approach to combine pattern recognition algorithms with semantic driven processes. The second approach, TCFS, used ontology based feature selection for clustering. Experiments proved that both techniques were efficient in clustering process, but the performance of TCFS was slightly better in terms clustering quality, but slow. In future, both these methods can be combined to take advantage of quality clustering in a fast manner.

References

1. Washio, T., Suzuki, E., Ting, K.M. and Inokuchi, A. (Eds.) "Advances in Knowledge Discovery and Data Mining, Proceedings of 12th Pacific-Asia Conference", PAKDD 2008 Osaka, Japan, Lecture Notes in Computer Science, 2008, Pp. 1-1102, ISBN: 978-3-540-68124-3.
2. Hafez, A.M. , "A Dynamic Approach for Knowledge Discovery of Web Access Patterns, ISMIS 2000", Lecture Notes In Computer Science; Vol. 1932 *Proceedings of the 12th International Symposium on Foundations of Intelligent Systems, Springer-Verlag* London, UK , 2000, Pp. 130-138.
3. Martin-Valdivia, M.T., Garcia-Vega, M. and Urena-Lopez, L.A. , "LVQ for text categorization using multilingual linguistic resource", Source: *Neurocomputing*, 2003, Vol. 55, Pp. 665-679.
4. Andrews, N.O. and Edward, A. (2007) Fox, Recent Developments in Document Clustering, http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf, Last Access Date : 24-03-2011.
5. Shawkat Ali, A.B.W. , " K-means Clustering Adopting rbf-Kernel, Data Mining and Knowledge Discovery Technologies", David Taniar (Ed.), 2008, Pp. 118-142.
6. Decherchi, S., Gastaldo, P., Redi, J. and Zunino, R. "K-Means Clustering for Content-Based Document Management in Intelligence, Advances in Artificial Intelligence for Privacy Protection and Security", *Intelligent Information Systems*, 2009, Vol. 1. Pp. 287-323.
7. Thangamani, M. and Thangaraj, P. , "Integrated Clustering and Feature Selection Scheme for Text Documents", *Journal of Computer Science*, 2010, Vol. 6, No.5,Pp. 536-541.
8. Raja,K. and Narayanan, C.P., "Clustering Technique with Feature Selection for Text Documents", *Proceedings of the Int. Conf. on Information Science and Applications* ICISA 2010, Pp.296--300.
9. Manning, C.D., Raghavan, P. and Schütze, H., " Introduction to Information Retrieval. Cambridge University Press", Cambridge, 2008.
10. Ridella, S., Rovetta, S. and Zunino, R., " Plastic algorithm for adaptive vector quantization", *Neural Computing and Applications*, 1998, Vol. 7, Pp. 37-51,
11. Dhillon, I.S. , "Co-clustering documents and words using bipartite spectral graph partitioning", *Knowledge Discovery and Data Mining*, 2001, Pp. 269–274.
12. Hand, D., Mannila, H. and Smyth, P., "Principles of Data Mining", MIT Press, Cambridge, MA, 2004.
13. Cai, D., He, X., and Han, J., " Document Clustering Using Locality Preserving Indexing", *IEEE Transaction on knowledge and data engineering*, 2005, Vol17, Pp. 1624-1637.
14. Shahnaz, F., Berry , M.W., Pauca, V.P. and Plemmons, R.J., " Document clustering using nonnegative matrix factorization", *Information Processing and Management* ,2006, Vol. 42, Pp. 373–386.
15. Sebastiani, F., "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, 2002, Vol. 34, No. 1, Pp. 55-59.
16. Girolami, M., "Mercer kernel based clustering in feature space", *IEEE Transactions on Neural Networks*, 2002, Vol.13, Pp. 2780-2784.