ISSN: 2321-7782 (Online)
Volume 2, Issue 2, February 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Paper / Case Study
Available online at: www.ijarcsms.com

Analysis on Clustering Techniques based on Similarity of Text Documents

B. Suganya[1]
PG Scholar
Department of CSE
SNS College of Engineering
India

P. Kirthika[2]
Assistant Professor
Department of CSE
SNS College of Engineering
India

Abstract: ***Text mining is the analysis of data contained in natural language text [6]. It is the process of extracting information from text. Text analysis involves information retrieval, lexical analysis, pattern recognition, information extraction. The main challenge in text mining is to find the similarity between documents in order to group the similar documents. The word frequency distributions are identified to find the similarity between various documents. The overarching goal is, essentially, to turn text into data for analysis. Vector space model was used to classify the relevant documents. Long documents are poorly represented because they have poor similarity values and keywords must precisely match the document terms. Documents might have similar context but different term vocabulary won't be associated which leads to less accuracy. Spectral based approach was used and it was suitable only to short queries.***

I. INTRODUCTION

Text mining is the field in data mining which is the process of extracting data from the large storage. [1] It is the process in knowledge discovery in database (KDD). Text indexing methods have been developed to handle unstructured documents. Traditional information retrieval technique became inadequate to retrieve documents from large storage database. Typically, only a small fraction of the many available documents will be relevant to a given individual user.

Information Retrieval

Retrieving the relevant information from various information resources. Efficient way to search relevant documents in web is to supply the query terms related to the information which you want to find. Search engine will then collect the necessary documents which are associated to the query terms provided. Information retrieval problem is to collect all the relevant documents based on the user's query from the database.

Information Extraction

It is the process of extracting structured information from unstructured or semi-structured documents. It refers to identify the words or feature terms within the document. It can be also described as content extraction out of images/audio/video could be seen as information extraction.

Text mining achieves this by applying techniques from natural language processing, grammatical analysis, speech tagging. Text mining has its application in market analysis, customer relationship management (CRM). To identify similarity between the various documents precision and recall are calculated. Precision is the measurement of deviation from true values and its scatter. Fraction of retrieved instance that is relevant. Recall is the fraction of relevant instances that are retrieved.

$$\text{Precision} = \frac{\text{No.Of correctly retrieved documents}}{\text{No. Of retrieved documents}}$$

$$\text{Recall} = \frac{\text{No. Of correctly retrieved documents}}{\text{No. Of documents in relevant category}}$$

To identify the topic of a document, the content of the document must be analysed. By obtaining a clear understanding of a page, it is possible to find more relevant documents. Vector space model has been the dominant method used to analyse the text in information retrieval system. This model represents every document as vector so that it can be compared with other document vectors. It compares the deviation of angle between each document vector and original query vector to find similarity amongst documents. Many information retrieval systems employs vector space model to classify text. Spectral based information retrieval method able to use both term count and position information to obtain high precision document ranking.

Text categorization

It refers assigning natural language texts to predefined categories based on their content. It supports document retrieval, extracting data from texts. This technique is keyword-based.

II. LITERATURE REVIEW

1. Distributional features

Distribution features are used to express the distribution of a word in a document [2]. Frequency of the word and position of the first appearance of the word are characterized as features. It uses TFIDF scheme to identify the features. Term frequency be the number of occurrences of term t in the document d, that is f(d,t). The term-frequency matrix TF(d,t) measures the association of the term t with respect to the given document d. Inverse document frequency is the number of documents which have the term t.

TF-IDF(d,t)=TF(d,t)*IDF(t)

kNN and SVM are used to test the effect of distributional features. The position of all the appearances of the word is identified and indexed. Distributional features are useful for text categorization and essential when the documents are long enough and informal.

Disadvantages:

- It is a theoretical approach
- Other methods can be used to improve the performance.

2. Vector Space Model

Vector space model is used to represent text documents as vectors [6]. It is used in information filtering, retrieval, indexing, relevancy ranking. Each text document is represented as vector. By comparing the angle of each document vector the similarity between those documents can be calculated.

$\text{Sim}(v1,v2) = v1.v2 / |v1||v2|$

If the term occurs in the document, its value in the vector is non-zero otherwise it is indicated as zero. Relevance ranking can be made for the documents using assumptions of document similarity theory.

Disadvantages:

- Spatial information contained in the documents may lose.
- Long documents are poorly represented.
- Not a promising way to discriminate contextual similarity.

3. Fourier domain scoring

Fourier domain scoring is the superior method because it makes use of the spatial information within the document rather than the count of each term [4]. This has given improvement in precision comparing to vector space model. It not only records number of times each word appears in the document but also the position of the words. Inverted index is created for the collected words. Pre-weighting is performed and perform fourier transform to calculate document scores. These informations are compared against other words. It shows high performance for short queries. DCT It is used to extract information from the text documents and it provides document ranking without having to store excessive data.FDS requires more disk space to store index. To reduce the storage cost DCT is used, which does not require more information. It provides high precision.

Disadvantages:

- Its index requires large storage space.
- Computational time increases.
- Still the performance can be improved.

4. Phrase-based document indexing

Document clustering technique mostly rely on single term analysis of document data set [5].The phrase similarity can be calculated using document index graph based on the list of matching phrases between the documents. DIG represents a directed graph G=(V,E). V denotes nodes in which each node represents a unique word in the document. E denotes edges and each edge is an ordered pair of nodes(*vi*,*vj*). If *vj* is successive to the node *vj* in any document then there will be an edge between *vi* to *vj*. Indexing information is stored in the graph nodes in the form of document table. When a new document is added, the graph is updated with the new sentence information. New words are added and connected with other nodes to reflect the sentence structure. Matching phrases that occur between the documents are calculated to find the similarity of the documents.

Disadvantages:

- With a very large document set, the graph could become more complex.

5. K- Means:

It randomly selects k objects each represents center. [6] For thr remaining objects, each object is assigned to the cluster to which it is the most similar. It is assigned based on the distance between the object and the cluster mean. Then new mean for each cluster is calculated. Cluster mean is updated to find the mean value of the cluster and it is repeated until no changes occurred. This is relatively scalable and efficient in processing large data sets. EM algorithm extends k means which assigns each object to each cluster according to a weight representing its probability of membership.

Disadvantages:

- It is not suitable for discovering clusters with different size.
- It is influenced by outliers.

6. k-Mediods:

It takes the actual objects to represent the clusters, using representative object for every cluster [6]. All remaining objects are clustered with the representative object to which it is most similar. Representative object is choosen arbitrarily and the representative object is replaced by non representative object until the quality of cluster is improved. PAM is k-mediods algorithm which tries to make the better choice of cluster representatives. All the pairs of objects are analysed and a object in that pair is considered as representative object. Quality of the cluster is calculated for each combination.

Disadvantages:

- Still the accuracy can be improved.

III. CONCLUSION

Distributional features are used to determine the word distribution but it is a theoretical approach. Vector space model does not identify semantic similarity among the documents. Fourier domain scoring is used to index the documents with high computational time. Phrase based document indexing method is complex for large document. To improve the accuracy in determining the similarity between various documents dimensionality techniques can be used.

References

1. G. Salton, M. McGill, Eds. Introduction to Modern Information Retrieval.New York: McGraw-Hill, 1983.
2. X. B. Xue and Z. H. Zhou, "Distributional features for text categorization,"IEEE Trans. Knowl. Data Eng., vol. 21, no. 3, pp. 428–442, Mar. 2009.
3. L. A. F. Park, M. Palaniswami, and K. Ramamohanarao, "A novel document ranking method using the discrete cosine transform," IEEE Trans.Pattern Anal. Mach. Intell., vol. 27, no. 1, pp. 130–135, Jan. 2005.
4. Park L A F, Ramamohanarao K and Palaniswami M, "Fourier domain scoring: A novel document ranking method," IEEE Trans. Knowl. Data Eng., vol. 16, no. 5, pp. 529–539, May 2004.
5. K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for web document clustering," IEEE Trans. Knowl. Data Eng., vol. 16, no. 10, pp. 1279–1296, Oct. 2004.
6. Jiawei Han and Micheline Kamber Data Mining: Concepts and Techniques,Second Edition.

