Clustering Technique using Concepts for Web Documents

Hmway Hmway Tar, Thi Thi Soe Nyunt
University of Computer Studies, Yangon
hmwaytar34@gmail.com, ttsoenyunt@gmail.com

Abstract

Web document clustering becomes an essential technology with the popularity of the Internet. That also means that fast and high-quality document clustering techniques play core topics. One of the main issues for clustering is the feature selection for the documents. The selected features should contain sufficient or more reliable information about original web documents. Feature selection is important because some of the irrelevant or redundant feature may misguide the clustering result. To counteract this issue, this paper proposes the concept weight for feature selection which can improve the efficiency and accuracy of clustering. The system is designed to perform document preprocessing, weight estimation and clustering process that uses the term weight and semantic weight. This paper introduces a method which proposed the concept weight for clustering process.

Key words: Clustering, feature selection, concept weight

1. Introduction

With the booming of the Internet, the World Wide Web currently contains billions of documents. This causes difficulty in finding the desired information by users. Many search engines come out to help users finding their desired information but all search engines still return hundreds of irrelevant web pages that do not fulfill the user's query. Many search engines use clustering to group documents that are relevant to the user's query before returning them to the user.

How to explore and utilize the huge amount of text documents is a major question in the areas of information retrieval and text mining methods that are developed to help users effectively navigate, summarize, and organize text documents become an important factor. By organizing a large amount of documents into a number of meaningful clusters, document clustering can be used to browse a collection of documents or to organize the results returned by a search engine in response to a user’s query. It can significantly improve the precision and recall in information retrieval systems, and it is an efficient way to find the nearest neighbors of a document. Document clustering is a fundamental operation used in unsupervised document organization, automatic topic extraction, and information retrieval. Clustering involves dividing a set of objects into a specified number of clusters. The major concept is therefore a collection of data instances which are similar to each other and are dissimilar to data instances in other clusters. There area two major clustering techniques: Partitioning and Hierarchical. Hierarchical techniques produce a nested sequence of partition, with a single, all inclusive cluster at the top and single clusters of individual points at the bottom. Partition techniques seek to partition a collection of documents into a set of non-overlapping groups, so as to maximize the evaluation value of clustering. Although the hierarchical clustering technique is often portrayed as a better quality clustering approach, the time complexity of this approach is quadratic.

In recent years, it has been recognized that the partitioned clustering approach is well suited for clustering a large document dataset to their relatively low computational requirements. The time complexity of it is almost linear, which makes it widely used.

The problem of document clustering is generally defined as follows: Given a set of documents, would like to partition them into a predetermined or an automatically derived number of clusters, such that the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters. In other words, the documents in one cluster share the same topic, and the documents in different clusters represent different topics.

In most existing document clustering algorithms, documents are represented using the vector space model which treats a document as a bag of words. A major characteristic is the high dimensionality of the feature space, which imposes a big challenge to the performance of clustering algorithms. They could not work efficiently in high dimensional feature spaces due to the inherent sparseness of the data. Another problem is that not all features are important for document clustering. Some of the features may be redundant or irrelevant. Some may even misguide the clustering result, especially when there are more irrelevant features than relevant ones. In such case, selecting a subset of original features often leads to better clustering performance. Feature selection not only reduces the high dimensionality of the feature space, but also provides better data understanding, which improves the clustering result. The selected feature set should contain sufficient or more reliable information about the original data set. For document clustering, this will be formulated into the problem of identifying the most informative words within a set of documents for clustering. This paper proposes a modify K-means clustering algorithm based on weight of the concept used for clustering process.

The rest of this paper is organized as follow. Section 2 discusses the related work. Section 3 illustrates the proposed system. Section 4 describes tf-idf model. Section 5 illustrates the similarity metric and in Section 6 describe k-mean algorithm for clustering. Finally, the paper is concluded with Section 8.

2. Related Work

Document clustering has been considered as one of the most crucial techniques for dealing with the diverse and large amount of information present on the World Wide Web. In particular, clustering is used to discover latent concepts in a collection of Web documents, which is inherently useful in organizing, summarizing, disambiguating, and searching through large document collections [7].

Many methods, including *k-means*, hierarchical clustering and nearest-neighbor clustering etc., select a set of key terms or phrases to organize the feature vectors corresponding to different documents. Suffix-tree clustering [5], a phrase-based approach, formed document clusters depending on the similarity between documents.

In [6], Feld proposed the *KDT* and *FACT* system to discover association rules based on keywords labeling the documents, the background knowledge of keywords and relationships between them. This is ineffective because a substantially large amount of background knowledge is required. Therefore, the use of term extraction modules have been propose to generate association rules by selected key words [8]. It is beneficial for us to obtain meaningful results without the need to label documents by human experts.

.

3. Overview of the Proposed System

The proposed system is designed to deal with three major phases. The document preprocessing, weight estimation and clustering process is done using the term weight and semantic weight. The document preprocessing is the initial phase for the system. Stop word elimination and stemming process are carried out in this phase. The term weights are computed using TF-IDF model. The term and its count are used in this model. The semantic weight estimation is done with the support of the ontology. Finally the clustering process is done using the term weight and semantic weight.

4. TF-IDF model

TF_IDF model is the most simple and sophisticated weighted schema used for feature extraction. Two major factors are worthily considered to determine the importance of a vocabulary in the document. The first one is the relative frequencies of occurrence with vocabulary in the document, called the Term Frequency (TF); the second is the frequencies of occurrence with document in all of document set are called the Document Frequency (DF). Term frequency measures the relative importance of a keyword in a document. If the value of the term frequency in the document (document in here pass by already remove that stop word) is very high, this term is very important that can represent this document. Document frequency measures how many documents have this term. The more the vocabulary appears in the document but in other documents few, the more this vocabulary is suitable for and used for distinguishing with other documents.

Synthesize the above two factors, we can calculate the importance of a vocabulary in a certain document by the product of term frequency and inverse of document frequency. This approach is called TFIDF; its formula lists as follows:

$$
\begin{aligned}
w_{i} &= tf_i \times idf \\
&= tf_i \times \log(\frac{n}{df_i}) \qquad (1)
\end{aligned}
$$

where w_i is the term weight of the term t_i in one document. The term weight value represents the significance of this term in a document. To calculate the term weight, the occurrence frequency of the term within a document and in the entire set of documents, tf_i is the number of occurrence of term i in the document.

5. The Similarity Metric

The similarity between two documents needs to be measured in a clustering analysis. Over the years, many prominent ways have been used to compute the similarity between documents $\boldsymbol{m_p}$ and $\boldsymbol{m_j}$. The most commonly used distance functions for clustering algorithm are the Euclidean distance, Manhattan (city block) distance and Cosine correlation measure. The commonly used similarity measure in document clustering is the cosine correlation measure, given by

$$\cos(m_p, m_j) = \frac{m_p m_j}{|m_p||m_j|} \tag{2}$$

where $m_p\ m_j$ denotes the dot-product of the two document vectors. |.| indicates the length of the vector.

6. K-Means Algorithm for Clustering

The K-means algorithm is very popular for solving the problem of clustering a data set into k clusters. If the data set contains n documents, d_1, d_2, . . . , d_n, then the clustering is the optimization process of grouping them into k clusters so that the global criterion function

$$\sum_{j=1}^{k} \sum_{i=1}^{n} f(d2, cenj) \tag{3}$$

is either minimized or maximized. Centroid represents the centroid of a cluster cj, for j = 1. . . k, and f (di, cenj) is the clustering criterion function for a document di and a centroid cenj. When the cosine function is used, each document is assigned to the cluster with the most similar centroid, and the global criterion function is maximized as a result. This optimization process is known as an NP complete problem and the K-means algorithm was proposed to provide an approximate solution. The steps of K-means are given as follows:

1. Select k initial cluster centroids
2. For each document of the whole data set, compute the clustering criterion function with each cluster centroid. Assign each document to its best choice.
3. Recalculate k centroids based on the documents assigned to them.
4. Repeat Steps 2 and 3 until convergence.

7. Conclusion

Document clustering is an important tool in web mining. Most clustering methods do not completely satisfy special requirements for web document clustering. The paper articulates the unique requirements of Web document clustering. This paper introduces a novel clustering process based on concept estimation which is very useful when the requirement for precision is high.

References

[1] Berkhin , P. , 2002. Survey of clustering data mining techniques. Accrue Software Research Paper.

[2] Dhillon. I, Mallela.S , and Kumar, R. 2002. Enhanced Word Clustering for Hierarchical Text Classification, In Proceedings of the 8th ACM SIGKDD, 191-200, Edmonton, Canada

[3] Ghosh 2002, Scalable Clustering Methods for Data Mining. In Nong Ye (Ed.) Handbook of Data Mining, Lawrence Erlbaum

[4] Jarvis, R.A. and Patrick, E.A. 1973. Clustering using a similarity measure based on shared nearest neighbors. IEEE Transactions on Computers, C-22, 11

[5] O. Zamir and O. Etzioni. Web document clustering:a feasibility demonstration. In *Proceedings of 19th international ACM SIGIR conference on research and development in information retrieval (SIGIR 98)*, pages 46–54, 1998

[6] R. Feldman, Y. Aumann, A. Amir, W. Kl´osgen, and A. Zilberstien. Text mining at the term level. In *Proceedings of 3rd International Conference on Knowledge Discovery, KDD-97*, pages 167–172, Newport Beach, CA, 1998.

[7] R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 2(1):1–15, 2000

[8] Zhao Y. and Karypis G., 2004. Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering, Machine Learning, 55 (3): pp. 311-331