

A brief study on document clustering

Ksh. Nareshkumar Singh

Dept of Computer Science

Manipur University

nareshksh2711@gmail.com

Dr. H. Mamata Devi

Dept. of Computer Science

Manipur University

mamata_dh@rediffmail.com

Dr. Anjana Kakoti Mahanta

Dept. Of Computer Science

Gauhati University

anjana@gauhati.ac.in

Dr. N Ajith Singh

B.Voc Information and

IT Services

South East Manipur

College

ajithex@gmail.com

Abstract:

With the advance in digital technology has led to increase the digital collections of data exponentially. These collected data are in different form such as unstructured text data on World Wide Web, sensor data, images, video, sound etc. Among these text data is most widely used and huge in amount. Analysing and processing the text data requires many data mining techniques. Clustering, one of them, is an unsupervised learning approach that groups the similar data objects together without using any human label data. In this paper, we have presented the study on document clustering methods, document representation and factor that effect on the clustering result.

Keywords: Document Clustering, Document Representation, Similarity Measure, Clustering Evaluation

I. INTRODUCTION

When the document is being analysed, the text of the document is the basic unit to be considered. So for document clustering, we mainly focussed on text data. Today's digital world, mainly text data is growing exponentially and expected to be 2.5 Exabyte per day. The massive amount of text data needs to cluster based on their features so that it can study and analyze in less time. Similarity measure between data objects are lies at the central part of document clustering. For organizing the massive amount of documents collection, we need to do classification and clustering method of data mining. To perform clustering method, there are three main factors involves - document representation method, similarity measure (based on distance or conceptual) and their applying algorithms. Similarity measures between the objects play a significant role in data

objects clustering. Actually, similarity measure is a real value function that assigns a value between $[0,1]$ to a pair of documents, depending upon the degree of similarity between them. That means similarity measure indicates how much degree of closeness or separation of target objects are. Choosing the right similarity measure for a given dataset is great challenge to achieve good results of clustering. Accuracy of clustering is also based on the effectiveness of the defined similarity measure.

Documents are needed to transform into vector representation so that clustering algorithm works on it. Actually clustering algorithm works on the vector, not on the original text in the document. In vector representation, documents are represented using high dimension vectors where each unique term in the vocabulary represents a dimension. Sometime, high dimension

vector requires dimension reduction in order to improve the computational efficiency in handling the massive data. Here, we are not considering the dimension reduction techniques.

There are many clustering algorithms and it can be categorized based on their cluster model. It is quite difficult to choose the most suitable algorithm for a particular problem without experimental. So we should choose the algorithm experimentally, unless we can't decide one algorithm prefer one another. It should be noted that an algorithm is best for one data set may be fail in some other data set.

The residual parts of the paper are organised as follows. In section II, we give the information about document clustering. Section III, mentions about document representation. In section IV, we will discuss the taxonomy of clustering methods. Section V considering about evaluation of clustering result. Section VI gives conclusion.

II. DOCUMENT CLUSTERING

Document clustering is the process of clubbing the documents into many clusters so that similar documents belong to the same cluster and dissimilar documents belong to different clusters. The grouping data in a cluster based on maximising the intra class cluster and minimising the inter class cluster. Clustering is included in unsupervised learning because it is learned by an observation and doesn't require human expert to assign the documents to the classes.

Clustering can be divided into four subgroups- flat (or exclusive), hierarchical, hard, and soft. Flat clustering creates a set of clusters horizontally without any explicit structure that would relate clusters to each other. In hierarchical clustering, partition the data objects hierarchically

where documents in lower level cluster may be membership of documents in corresponding higher level cluster. In hard clustering, each document belongs to mutually exclusive clusters whereas in soft clustering, document can belong to more than one cluster.

Now, why are we interested in document clustering? Here, some of reasons are:

- Finding the similar documents
- Organising the large collection of documents
- Recommendation System
- Search Optimization

III. DOCUMENT REPRESENTATION

Document representation is required before doing document clustering because clustering algorithm works on vectors, not on the original text in the document. Hence, documents are often represented using high dimension vectors where each term in the vocabulary of the collection represents a dimension. There are several document representation models. Some of them are mentioned here:

- Bag-of-Words Model: In this model, documents are represented as bag of its word, ignoring the grammatical rule and word order. For example, a document 'Tom can run faster than Harry' and another document, 'Harry can run faster than Tom' are identical in this model view. It is simply collecting all the words in the document. This bag-of-word model is a special case of n-grams representation where $n=1$.
- n-gram representation: an n-gram is a contiguous sequence of n

words (characters) just like moving a window of n boxes on the document body. This n -gram representation can store some spatial information within the words. When $n=2$, it is called bigram model and $n=3$ is called trigram model and so on.

- **Vector Space Model (VSM) :** In VSM, documents are represented as vectors in an n - dimensional space, where n is the number of unique terms in the vocabulary. The weight value of each term can be computed by using different weighted schemes. The most commonly use weighting scheme is TF*IDF scheme. The vector representation doesn't consider the ordering of words in a document and semantic relations of words. That is, terms represented in the vector space are assumed to be mutually independent. Although, it is widely used due to its easy to calculate the similarity between documents, simple and useful for describing document features.
- **Word2vec Model:** Word2vec model is used for learning vector representation of words called "word embedding". Word2vec takes as its input a large text dataset and produce a vector space, with each unique word in the text dataset. Word vectors are located closely to one another in the vector space when words that share the common contexts in the text dataset. Word2vec can use either of

two model to produce the representation of words- Continuous Bag-of-Words (CBOW) model or Continuous Skip-gram model. CBOW determine a word from its context and Skip-gram determines the context from the word.

- **Dependency Graph:** Document can be represented as dependency graph G . This document dependency graph G is denote as $G=(V,E)$ where $V=\{v_1, v_2, \dots, v_n\}$ is the set of vertices where each vertices v_i represent the words w_i of the document and $E=\{e_1, e_2, \dots, e_n\}$ is the set of edges where each edge e_i between the vertices indicates that there is some relationship between words of the document. Each edge can be drawn without crossing over the words. Parsing phase is also required to construct dependency graph. The output from dependency graph is considered as tree, that explains how sentence connect the words to each other.

IV. CLUSTERING METHODS

How to do document clustering? Clustering can be performed by using various methods. In general, clustering methods can be divided into following categories as shown in figure 1.

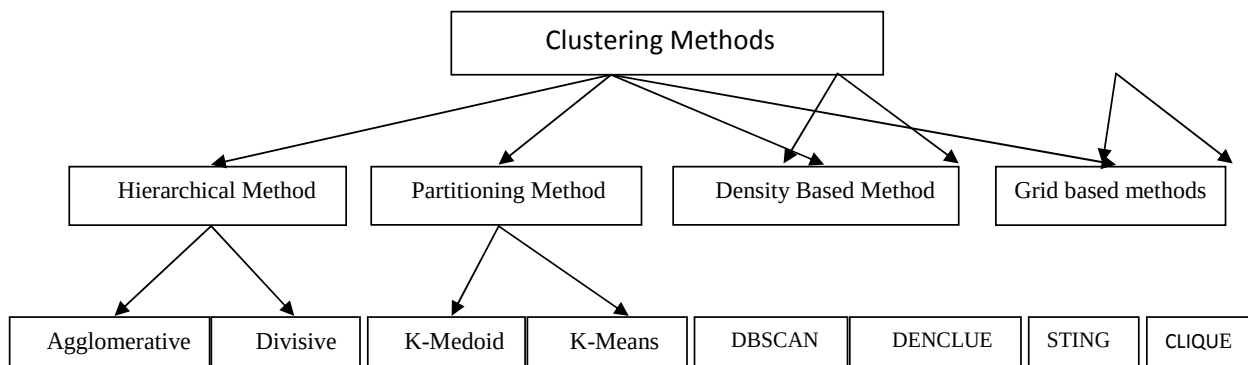


Fig1: Taxonomy of Clustering Methods

structure is formed by decomposing the given set of data objects. Hierarchical decomposition is performed either one of the following two methods

- **Agglomerative:** it is the bottom up approach where it starts by taking each individual object creates its own cluster and for every pair of clusters, value of disparity is measured in term of distance. At every step, the clusters with minimal distance are merged together. This process will continue iteratively until all the objects meet certain termination condition or are in a single cluster. Single cluster may become the hierarchy's root or may have two or more clusters but each cluster should have at least one object.

successive iteration, each cluster is split into smaller cluster. This process will continue until each object in one cluster or hold the termination condition.

ii) **Partitioning Method:** this method performs partition on the given data set, where each partition represents a cluster. Most of the partitioning methods are distance-based.

- **K-means:** this algorithm take mean or centroid of the objects in a cluster as cluster center. The main objective of this algorithm is to minimize the average squared distance of objects from their cluster centers.
- **K-medoids:** take an actual object in the cluster as a reference point, that point represents the cluster. Each objects assigned to the clusters of which the most similarity to the representative object. Its aim is to minimize the sum of the difference between each object in the cluster and its corresponding reference point.

iii) Density Based Method: grouping the data objects based on the concept of density so it can construct the cluster of arbitrary shape rather than spherical-shaped cluster.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): it finds the core objects which have dense neighbourhoods. To form a dense region as a cluster by connecting the core objects and their neighbourhoods.
- DENCLUE (DENSity-based CLUstering): A local maximum of the estimated density function is used to define the cluster. Data points that have same local maximum are put in the same cluster. Basically it is depend of a set of density distribution functions.

iv) Grid Based Method: the object space is subdivided into a finite number of cell that form a grid structure. All the clustering operations are performed on the grid structure.

- STING (STatistical INformation Grid): it is a grid based multiresolution clustering technique.
- CLIQUE (CLustering In QUEst): this grid-based method is used for finding density based clusters in subspaces.

A. Similarity or Distance measures in Algorithmic methods

Similarity or distance measures play a significant role in design of clustering

methods. Some methods measure the similarity between objects based on the distance, some other may be based on density or continuity, and other may be based on conceptual. Clustering can be performed in many ways based on different characteristics of the objects. In geometrical distance, two or more objects are belongs to the same cluster if they are closed according to the given distance. This is called *distance-based* clustering. In the case of *conceptual clustering*, objects are grouped in the same cluster according to their fit to descriptive concepts. For example, Are “car” and “horse” similar or are they in the same cluster? They are not similar based on their physical properties. But if we look at them functionally, a car and a horse both can be used as transportation tools. So in that sense, they are similar and belong to the same cluster. Thus, it really depends on our perspective, the way to look at the objects. So, in order to make clustering problem well defined, a user must define the perspective for assessing similarity. There are many similarity or distance measure methods namely- Euclidean distance, Cosine similarity, Jaccard Co-efficient relation etc. These measure methods are used according to the features of the data set which we have to use.

V. EVALUATION OF DOCUMENT CLUSTERING

Evaluation Measures: It is very difficult to evaluate the clustering result just like clustering itself. Two popular approach are there- internal evaluation and external evaluation.

Internal Evaluation: data that was clustered itself are used to evaluate the clustering result. That means we evaluate the clustering result using quantities and

features inherent in the data set. The optimal number of cluster is determined based on the internal validity index.

External Evaluation: measure how much close the clustering results is to the predetermined benchmark classes which are created by the human expert. Actually, external evaluation of clustering result based on data that was not used in the clustering. Some of measure of quality of clustering result using external criterion are:

- Entropy gives the information about distribution of documents from the various classes within each cluster. Smaller the entropy value, better the clustering quality.
- Purity is the total number of objects that are classified correctly.
- Precision of an object i tells that how many other objects belongs to the same category as the object i in the same cluster
- Recall tells that how many objects of the same category are belongs to the same cluster.
- F-measure takes both Precision (P) and Recall (R). It is defined as the harmonic mean of precision and recall.

$$F = \frac{2PR}{P+R}$$

VI. CONCLUSIONS:

The quality of document clustering depends on the following factors:

i) selection of the document representation method

ii) define a measure of proximity between data objects:

- for dense data such as time series or two-dimensional points, euclidean distance is more useful than others

- In case of sparse data like documents, or binary data, cosine similarity and jaccard measures are useful.
- The similarity or distance measures (namely Euclidean Distance, Manhattan Distance, Supremum distance, Cosine Similarity etc.) are appropriate for continuous variables. However, for binary variables a different approach is used namely Simple Matching Coefficients, Jaccard Coefficients, Hamming distance etc.

iii) choosing suitable algorithms based on nature of data objects

iv) an assessment of an output is needed to investigate the validity of the results.

REFERENCES

- [1] Ksh. Nareshkumar Singh, H. Mamata Devi, Anjana Kakoti Mahanta "Document representation techniques and their effect on the document Clustering and Classification: A Review", International Journal of Advanced Research in Computer Science(IJARCS), Volume 8, No. 5, May-June 2017
- [2] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques" Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA
- [3] Anna Huang, "Similarity Measures for Text Document Clustering" NZCSRSC, Christchurch, New Zealand, April 2008.
- [4] Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques", KDD Workshop on Text Mining, 2000
- [5] Jitendra Agrawal and Shikha Agrawal, "Applications of Population Based Algorithms for Document Clustering, an article in CSI, June 2016.

- [6] Evangelos E. Milios, M. Mahdi Shafiei, Singer Wang, Roger Zhang, Bin Tang and Jane, "A systematic study on document representation and dimensionality reduction for text clustering," Technical Report CS-2006-05, Faculty of Computer Science, Dalhousie University, July 11, 2006.
- [7] Zakaria Elberrichi and Karima Abidi, "Arabic text categorization: a comparative study of different representation modes," The International Arab Journal of Information Technology, Vol.9, No. 5, September 2012.
- [8] Wang, Y., Ni, X., Sun, J., Tong, Y. and Chen, Z; 2011. Representing Document as Dependency Graph for Document Clustering. Proc. 20th ACM international conference on Information and knowledge management, 2177-2180.
- [9] T. Mikolov, K.Chen, G. Corrado, J.Dean, "Efficient Estimation of Word Representations in Vector Space", arXiv:1301.3781v3 [cs.CL] 7Sep 2013
- [10] Quoc Le, Tomas Mikolov, "Distributed Representations of Sentences and Documents" Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196, May 16, 2014.
- [11] Ramkumar.S, Elakkiya.A, Emayavaramban.G, "Data Transfer Model - Tracking and Identification of Data Files Using Clustering Algorithms", International Journal of Latest Technology in Engineering, Management & Applied Science, Vol.3 (8), pp. 13-21, August 2014.
- [12] Elakkiya. A, Ramkumar. S, Suganya. M," A Survey on Different Clustering Algorithms and its Techniques", International Journal of Innovation and Scientific Research, Vol.10(2), pp.443-445, Oct-2014
- [13] S.Ramkumar, K.Sathesh Kumar," A Personalized Scheme For Incomplete And Duplicate Information Handling In Relational Databases", International Journal Of Engineering Sciences & Research Technology, Vol. 5(9), pp.360-369, Sep-2016.
- [14] https://en.wikipedia.org/wiki/Cluster_analysis

