IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

ISSUES ON TRADITIONAL AND MODERN TEXTUAL DOCUMENT CLUSTERING ALGORITHMS

Wael M.S. Yafooz*
[*] Department of computer science Faculty of Computer and Information Technology Al-Madinah International University (MEDIU) 40100 Shah Alam, Selangor, Malaysia.

DOI: 10.5281/zenodo.167097

ABSTRACT

The amount of digital data utilized in daily life has increased owing to the high dependence on such data. Most data can be stored in textual documents. With the rapid increase in the number of textual documents, users face problems in obtaining useful information. Thus, a method by which to manage data is required to give users an idea about content. In addition, techniques to increase the ratio of precision in information retrieval results are also needed. Therefore, the textual document clustering area is developed to represent the data in meaningful clusters. The two main factors encountered in the process of textual document clustering are efficiency and goodness or quality of data clusters. Efforts have been exerted to deal with these factors. These attempts can be categorized into either traditional or modern approaches. However, these attempts also face numerous issues. In this paper, we present the previous and current issues faced by textual document clustering algorithms to help text domain researchers understand these issues. This study provides researchers and students an overview about textual document clustering algorithms. Furthermore, this study can encourage researchers to find solutions to these issues.

KEYWORDS: Textual document clustering, frequent-term, partitional clustering, hierarchical clustering.

INTRODUCTION

Daily tasks have recently been converted to digital information, which is easy to retrieve. Most information is stored in textual documents. Such information is known as unstructured information, which is the most difficult to organize because of the requirement for other tools, such as news articles[1], personal documents, and discussion forums. By contrast, structured information is easy to understand and deal with it. Thus, the user can obtain such information through relational databases[2].

Textual document clustering (known as text clustering) is a technique for managing and organizing textual documents[3]. The clustering process, which is known as unsupervised learning, generally groups data objects based on the similarity between their attributes [4]. By contrast, classifications categorized objects into predefined classes. The clustering process is known as supervised learning. In clustering process textual documents that have similarities in content are grouped in the same data cluster. Textual documents that do not have any similarity are placed in different data clusters.

There are many attempts to cluster textual documents that can be categorized into traditional [5, 6] and modern textual clustering algorithms [7-12]. Traditional clustering algorithms have two main approaches: partitional and hierarchical document clustering. Partitional document clustering represents a textual document in a one-level view. The most popular algorithms in this category are k-mean and its variants. These algorithms can efficiently produce data clusters. However, the accuracy of data is insufficient compared with hierarchical methods, where a textual document is represented in a multi-level view that considers the topic and sub-topic. Although hierarchical clustering produces accurate data cluster with good quality, this method is time consuming. Both approaches encounter many issues, are discussed in the next section. Moreover, modern textual document clustering algorithms are introduced to mitigate the issues of traditional methods. This paper, present an over views of textual documents approaches and highlights the previous and current issues of textual document clustering.

This paper is organized as follows: Section 2 presents traditional and modern clustering algorithms. In section 3, the issues of traditional and modern clustering algorithms are discussed. Finally, the conclusion of this paper is in Section 4.

TRADITIONAL AND MODERN CLUSTERING ALGORITHMS

In ttraditional methods, there are two common approaches are partitional and hierarchical.The partitional methods are the most efficient (fast executing time) in the clustering process[6]. The most well-known algorithm is the k-mean clustering algorithm. This algorithm begin is initiated with a user parameter to enter the number of data clusters, which known as k-mean[13]. K-mean begins with a random initial data point (textual document) called seed, and the number of initial points is equal to the number of request clusters. The mean is then calculated to determine the centriod of data clusters. The distance between the centriod and data point is measured. The data points are grouped to the closest centriod based on the number of iterations. However, the data quality is poor. Thus, variations of k-mean, such as the bisecting k-mean and medios, have been introduced to overcome these weaknesses[6]. The algorithms in this category generally do not represent a textual document in an understandable form, unlike hierarchical clustering [10].

The two types of hierarchical clustering algorithms are agglomerative and division. Division can be characterized as top-bottom, where as agglomerative is bottom–up. Agglomerative algorithms, Unweighted Pair Group Method with Arithmetic Mean (UPMGA), are the most suitable for distance measure for textual document representation. UPMGA can efficiently produce quality data clusters [3, 6]. Nonetheless, a hierarchical cluster is more suitable or representing a textual document cluster than a partitional approach because the former is based on a multi-level concept that focuses on a topic and sub topic style, whereas the latter represents clusters in a single level. The bisecting k-mean which is mix of partitonal and hierarchical is outperform both in producing a good quality of data clusters [6]. To mitigate previous issues, which are discussed in Section 3, modern methods have been proposed. Modern methods focus on efficiency by reducing the number of words and improving the quality of clustering based on semantic similarity using the WorldNet database [14] or Wikipedia [15]. Nevertheless, both textual document clustering approaches confront many issues.

ISSUES OF TEXTUAL CLUSTERING ALGORITHMS

Textual clustering algorithms, both traditional and modern, are confronted by numerous issues, including high dimensionality of data, scalability, accuracy, overlapping, predefined number of clusters, and cluster representation.

High Dimensionality of Data

High dimensionality of data occurs when the clustering algorithm uses all the words found in the corpus or collection of textual data. This issue is commonly confronted by traditional methods, such as k-mean and hierarchical. In such methods, the textual document is considered as a bag of words using the vector space model for weighing the term frequency and distance measures, such as cosine similarity. This problem is partially solved by modern methods that select only the most frequent or common words. Modern methods use document preprocessing to clean and prepare the textual document before the clustering process. Clustering then uses frequent words using cosine similarity or a new similarity measure. However, if the collection of documents is large, the selected words will be numerous.

Scalability

Traditional algorithms suffer when applied to a large dataset, but such algorithms can run well on a small dataset. Modern methods face the same problem. Although a few studies focused on incremental clustering[16, 17], they fail to consider all terms in textual documents and therefore yield textual data clusters with poor quality.

Accuracy

Accuracy is also known as the goodness of data clusters. Traditional algorithms deal with the words in textual documents as if no semantic relation or meaning exists among the words. Although modern approaches focus on semantic clustering for textual documents using WorldNet [14] or Wikipedia as a repository of knowledge, the modern research work focused on concept or statement structure. Another way to produce high-quality data clusters is to use a named entity [11, 18] or ontology.

Overlapping

Overlapping issues occur when the document belongs to more than one cluster, which is known as soft clustering. Such type of clustering is available in partitional clustering algorithms. Although hierarchical clustering algorithms are constructed using the topic-subtopic style, such algorithms support hard clustering, where textual documents belong to only one cluster. Some works used concepts of disjoint clusters before or after the clustering process.

Predefined Number of Clusters

All traditional methods require users to enter the number of clusters as an input parameter to determine the number of textual data clusters. Such methods require the user to have prior knowledge about the content of corpus (collection of documents). In modern methods, the number of clusters is determined automatically. Some research also used a mix of traditional and modern methods.

Cluster Representation

Cluster representation is a method of viewing a collection of textual documents. As previously mentioned, traditional methods can be partitional or hierarchical. Modern methods focus on hierarchical multi-level representation. Certainly, hierarchical representation is preferable for textual documents.

MODERN DOCUMENT CLUSTERING TECHNIQUES

There are many documents clustering algorithms. These algorithms can be categorized into two approaches: *Classical and Modern Document Clustering Approaches*. Modern document clustering techniques are methods used to perform textual document clustering. These methods can be classify into three categories are; Term frequent, semantic-based and named entity. Both approaches shown in Table 1.

Table 1: Comparison between Traditional and Modern Textual Clustering Algorithm

Approach	Traditional		Modern		
Method	Partiti-onal	Hierarchical	Frequent Term	Semantic	Named Entity
Performanc e	Poor	Good	Very Good	Very Good	Very Good
Accuracy	Good	Poor	Very Good	Very Good	Very Good
Terms	ALL	ALL	Frequent	Frequent	Named Entity
Semantic	No	No	No	Yes	No
Literature	[5, 6]		[8, 10-13, 19, 20]		

Table 1 demonstrates the comparison between the traditional and modern textual document algorithms. It shows that the traditional methods rely on a bag-of-words strategy by weighing the term frequency of words. Such methods use a distance measure, such as cosine similarity. While Modern approaches are different in that such approaches treat textual documents as a bag of semantic meanings or concepts. The efficiency of algorithms and accuracy of data clusters achieved using modern methods are better compared with that achieved using traditional methods.

CONCLUSION

Textual document clustering is an interesting technique for many research areas, particularly in handling a massive unstructured text data. In this paper, we highlighted the most important issues confronting textual document clustering algorithms. In previous studies and applications, partitional or hierarchical clustering algorithms were used. These algorithms involve numerous issues such as data high dimensionality and low efficiency and accuracy. Some of these issues were partially solved when modern algorithms were used. However, the accuracy of data clusters remains as challenge in the textual document clustering process.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

REFERENCES

1. Yafooz, Wael M.S. , Siti ZZ Abidin, and Nasiroh Omar, *Challenges and issues on online news management.* Control System, Computing and Engineering (ICCSCE),IEEE International Conference on., 2011.
2. Yafooz, Wael M.S., Siti ZZ Abidin, Nasiroh Omar, and Zanariah Idrus. *Managing unstructured data in relational databases*. in *Systems, Process & Control (ICSPC), 2013 IEEE Conference on*. 2013: IEEE.
3. Luo, Congnan, Yanjun Li, and Soon M. Chung, *Text document clustering based on neighbors*. Data & Knowledge Engineering 2009. **68**: p. 1271–1288.
4. Jain, Anil.K., Narasimha Murty, and Patrick J. Flynn, *Data Clustering: A Review*. ACM computing surveys (CSUR), 1999. **31.3**: p. 264-323.

5.Cutting, Douglass R., David R. Karger, Jan O. Pedersen, and John W. Tukey, *Scatter/gather: A cluster-based approach to browsing large document collections.* 15th annual international ACM SIGIR conference on Research and development in information retrieval., 1992.

6. Steinbach, Michael, George Karypis, and Vipin Kumar, *A Comparison of Document Clustering Techniques.* KDD workshop on text mining, 2000. **Vol. 400**.
7. Fung, Benjamin C.M., Ke Wangy, and Martin Ester, *Hierarchical Document Clustering Using Frequent Itemsets.* Proceedings of the SIAM international conference on data mining, 2003. **30. No. 5**.
8. Beil, Florian, Martin Ester, and Xiaowei Xu, *Frequent Term-Based Text Clustering.* Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. , 2002.
9. Malik, Hassan H. and John R. Kender, *High Quality, Efficient Hierarchical Document Clustering using Closed Interesting Itemsets.* Data Mining, ICDM'06. Sixth International Conference on. IEEE, 2006.
10. Li, Yanjun, Soon M. Chung, and John D. Holt, *Text document clustering based on frequent word meaning sequences.* Data & Knowledge Engineering, 2008. **64.1**: p. 381-404.
11. Montalvo, Soto, Fresno Víctor, and Martínez. Raquel, *NESM: a Named Entity based Proximity Measure for Multilingual News Clustering.*Procesamiento de Lenguaje Natural, 2012. **48**: p. 81-88.
12. Zhang, Wen, Taketoshi Yoshida, Xijin Tangc, and Qing Wanga, *Text clustering using frequent itemsets.* Knowledge-Based Systems, 2010. **23.5**: p. 379-388.
13. Zhao, Ying and George Karypis, *Evaluation of hierarchical clustering algorithms for document datasets.* Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.
14. Miller, George, *WordNet: A Lexical Database for English.* Communications of the ACM 1995. **38.11**: p. 39-41.
15. Milne, David and Ian H. Witten, *Learning to Link with Wikipedia.* Proceedings of the 17th ACM conference on Information and knowledge management. ACM,, 2008.
16. Hammouda, Khaled M. and Mohamed S. Kamel, *Incremental Document Clustering Using Cluster Similarity Histograms.* Web Intelligence, Proceedings. IEEE/WIC International Conference on. IEEE, 2003.
17. Gil-García, Reynaldo and Aurora Pons-Porrata, *Dynamic hierarchical algorithms for document clustering.* Pattern Recognition Letters 2010. **31** p. 469–477.
18. Montalvo, Soto, Raquel Martnez, Arantza Casillas, and Victor Fresno, *Bilingual News Clustering Using Named Entities and Fuzzy Similarity.* Text, Speech and Dialogue. Springer Berlin Heidelberg, 2007.
19. Shehata, Shady, Fakhri Karray, and Mohamed Kamel, *Efficient Concept-Based Mining Model for Enhancing Text Clustering.* Knowledge and Data Engineering, IEEE Transactions on, 2010.
20. Yafooz, Wael M.S., Siti ZZ Abidin, Nasiroh Omar, and Rosenah A Halim. *Dynamic semantic textual document clustering using frequent terms and named entity*. in *System Engineering and Technology (ICSET), 2013 IEEE 3rd International Conference on*. 2013: IEEE.