

# A COMPREHENSIVE SURVEY OF THE EXISTING TEXT CLUSTERING TECHNIQUES

<sup>1</sup>Patil Pravin Ishwar, <sup>2</sup>Prof. Gajendra Singh

**ABSTRACT:** Clustering analysis is a vibrant field of research intelligence and data mining. The fundamental concept of clustering analysis lies in using its salient characteristics to calculate the degree of similar relationships among objects and attaining automatic classification. In the recent years, clustering of data in large data sets has become a rallying area of research and it is gaining momentum. Clustering of document is important for the purpose of document organization, summarization, topic extraction and information retrieval in an efficient way. Initially, clustering is applied for enhancing the information retrieval techniques. Of late, clustering techniques have been applied in the areas which involve browsing the gathered data or in categorizing the outcome provided by the search engines for the reply to the query raised by the users. In this paper, we are providing a comprehensive survey over the document clustering.

## INTRODUCTION

Document clustering is also applicable in producing the hierarchical grouping of document (Ward 1963). In order to search and retrieve then information efficiently in Document Management Systems (DMS), the metadata set should be created for the documents with adequate details. But just one metadata set is not enough for the whole document management systems. This is because various document types need different attributes to be distinguished appropriately.

So clustering of documents is an automatic grouping of text documents into clusters such that documents within a cluster have high resemblance in comparison to one another, but are different from documents in other clusters. Hierarchical document clustering (Murtagh 1984) categorizes clusters into a tree or a hierarchy that benefits browsing.

Information Retrieval (IR) (Baeza 1992) is the field of computer science that focuses on the processing of documents in such a way that the document can be quickly retrieved based on keywords specified in a user's query. IR technology is the foundation of web-based search engines and plays a key role in biomedical research, as it is the basis of software that aids literature search.

## LITERATURE SURVEY

Document clustering is the process of categorizing text document into a systematic cluster or group, such that the documents in the same cluster are similar whereas the documents in the other clusters are dissimilar. It is one of the vital processes in text mining. Liping (2010) emphasized that the expansion of internet and computational processes has paved the way for various clustering techniques. Text mining especially has gained a lot of importance and it demands various tasks such as production of granular taxonomies, document summarization etc., for developing a higher quality information from text.

Likas *et al.* (2003) proposed the global K-means clustering technique that creates initial centers by recursively dividing data space into disjointed subspaces using the K-dimensional tree approach. The cutting hyper plane used in this approach is the plane that is perpendicular to the maximum variance axis derived by Principal Component Analysis (PCA). Partitioning was carried out as far as each of the leaf nodes possess less than a predefined number of data instances or the predefined number of buckets has been generated. The initial center for K-means is the centroids of data that are present in the final buckets. Shehroz Khan and Amir Ahmad (2004) stipulated iterative clustering techniques to calculate initial cluster centers for K-means. This process is feasible for clustering techniques for continuous data.

Agrawal *et al.* (2005) ascribed data mining applications and their various requirements on clustering techniques. The main requirements considered are their ability to identify clusters embedded in subspaces. The subspaces contain high dimensional data and scalability. They also consist of the comprehensible ability of results by end-users and distribution of unpredictable data transfer.

The main limitation of K-means approach is that it generates empty clusters based on initial center vectors. However, this drawback does not cause any significant problem for static execution of K-means algorithm and the problem can be overcome by executing K-means algorithm for a number of times. However, in a few applications, the cluster issue poses problems of erratic behavior of the system and affects the overall performance. Malay Pakhira *et al.* (2009) mooted a modified version of the K-means algorithm that effectively eradicates this empty cluster problem. In fact, in the experiments done in this regard, this algorithm showed better performance than that of traditional methods.

Uncertainty heterogeneous data streams (Charu Aggarwal *et al.* 2003) are seen in most of the applications. But the clustering quality of the existing approaches for clustering heterogeneous data streams with uncertainty is not satisfactory. Guo-Yan Huang *et al.* (2010) posited an approach for clustering heterogeneous data streams with uncertainty. A frequency histogram using H-UCF helps to trace characteristic categorical statistic. Initially, creating 'n' clusters by a K-prototype algorithm, the new approach proves to be more useful than UMicro in regard to clustering quality.

Alam *et al.* (2010) designed a novel clustering algorithm by blending partitional and hierarchical clustering called HPSO. It utilized the swarm intelligence of ants in a decentralized environment. This algorithm proved to be very effective as it performed clustering in a hierarchical manner.

Shin-Jye Lee *et al.* (2010) suggested clustering-based method to identify the fuzzy system. To initiate the task, it tried to present a modular approach, based on hybrid clustering technique. Next, finding the number and location of clusters seemed the primary concerns for evolving such a model. So, taking input, output, generalization and specialization, a HCA has been designed. This three-part input-output clustering algorithm adopts several clustering characteristics simultaneously to identify the problem.

Only a few researchers have focused attention on partitioning categorical data in an incremental mode. Designing an incremental clustering for categorical data is a vital issue. Li Taoying *et al.* (2010) lent support to an incremental clustering for categorical data using clustering ensemble. They initially reduced redundant attributes if required, and then made use of true values of different attributes to form clustering memberships.

Crescenzi *et al.* (2004) cited an approach that automatically extracts data from large data-intensive web sites. The “data grabber” investigates a large web site and infers a scheme for it, describing it as a directed graph with nodes. It describes classes of structurally similar pages and arcs representing links between these pages. After locating the classes of interest, a library of wrappers can be created, one per class with the help of an external wrapper generator and in this way suitable data can be extracted.

Miha Grcar *et al.* (2008) mulled over a technique about the lack of software mining technique, which is a process of extracting knowledge out of source code. They presented a software mining mission with an integration of text mining and link study technique. This technique is concerned with the inter links between instances. Retrieval and knowledge based approaches are the two main tasks used in constructing a tool for software component. An ontology-learning framework named LATINO was developed by Grcar *et al.* (2006). LATINO, an open source purpose data mining platform, offers text mining, link analysis, machine learning, etc.

Similarity-based approach and model-based approaches (Meila and Heckerman 2001) are the two major categories of clustering approaches and these have been described by Pallav Roxy and Durga Toshniwal (2009). The former, capable of maximizing average similarities within clusters and minimizing the same among clusters, is a pairwise similarity clustering approach. The latter tries to generate techniques from the documents, each approach representing one document group in particular.

Document clustering is becoming more and more important with the abundance of text documents available through World Wide Web and corporate document management systems. But there are still some major drawbacks in the existing text clustering techniques that greatly affect their practical applicability. The drawbacks in the existing clustering approaches are listed below:

- Text clustering that yields a clear cut output has got to be the most favorable. However, documents can be regarded differently by people with different needs vis-à-vis the clustering of texts. For example, a businessman looks at business documents not in the same way as a technologist sees them (Macskassy *et al.* 1998). So clustering tasks depend on intrinsic parameters that make way for a diversity of views.
- Text clustering is a clustering task in a high-dimensional space, where each word is seen as an important attribute for a text. Empirical and mathematical analysis have revealed that clustering in high-dimensional spaces is very complex, as every data point is likely to have the same distance from all the other data points (Beyer *et al.* 1999).
- Text clustering is often useless, unless it is integrated with reason for particular texts are grouped into a particular cluster. It means that one output preferred from clustering in practical settings is the explanation why a particular cluster result was created rather than the result itself. One usual technique for producing explanations is the learning of rules based on the cluster results. But this technique suffers from a high number of features chosen for computing clusters.

## CONCLUSION

As clustering plays a very vital role in various applications, many researches are still being done. The upcoming innovations are mainly due to the properties and the characteristics of existing methods. These existing approaches form the basis for the various innovations in the field of clustering. From the existing clustering techniques, it is clearly observed that the clustering techniques based on GA, fuzzy and ontology provide significant results and performance. Hence, this research concentrates mainly on the semantic clustering based on GA, NDRGA, ACO and fuzzy ontology clustering for better performance.

## REFERENCES

- [1] Agrawal, Rakesh, Gehrke, Johannes, Gunopulos, Dimitrios, Raghavan and Prabhakar, “Automatic subspace clustering of high dimensional data”, *Data Mining and Knowledge Discovery* (Springer Netherlands) Vol. 11, pp. 5-33, DOI:10.1007/s10618-005-1396-1, 2005.
- [2] Alam, S., Dobbie, G., Riddle, P. and Naeem, M.A. “Particle Swarm Optimization Based Hierarchical Agglomerative Clustering”, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 2, pp. 64-68, 2010.

- [3] Baeza-Yates, R.A. "Introduction to Data Structures and Algorithms Related to Information Retrieval", In Information Retrieval: Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice- Hall, Inc., Upper Saddle River, New Jersey, pp. 13-27, 1992.
- [4] Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu, "A Framework for Clustering Evolving Data Streams", Proceedings of the 29th international conference on Very Large Data Bases (VLDB), pp. 81-92, 2003.
- [5] Crescenzi valter, Giansalvatore Mecca, Paolo Merialdo and Paolo Missier, "An Automatic Data Grabber for Large Web Sites", VLDB , pp. 1321-1324, 2004
- [6] Grcar, M., Mladenic, D., Grobelnik, M., Fortuna, B. and Brank, J. "Ontology Learning Implementation", Project report IST-2004-026460 TAO, WP 2, D2.2, 2006.
- [7] Guo-Yan Huang, Da-Peng Liang, Chang-Zhen Hu and Jia-Dong Ren, "An algorithm for clustering heterogeneous data streams with uncertainty", 2010 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 4, pp. 2059-2064, 2010.
- [8] Li Taoying, Chne Yan, Qu Lili and Mu Xiangwei, "Incremental clustering for categorical data using clustering ensemble", 29th Chinese Control Conference (CCC), pp. 2519-2524, 2010.
- [9] Likas, A., Vlassis, N. and Verbeek, J.J. "The Global k-means Clustering algorithm", Pattern Recognition , Vol. 36, No. 2, pp. 451-461, 2003.
- [10] Lijuan Jiao and Liping Feng, "Text Classification Based on Ant Colony Optimization", Third International Conference on Information and Computing (ICIC), Vol. 3, pp.229 - 232, 2010.
- [11] Macskassy, S.A., Banerjee, A. Davison, B.D. and Hirsh, H. "Human Performance On Clustering Web Pages: A Preliminary Study", In Proc. of KDD-1998, New York, USA, pp. 264-268, Menlo Park, CA, USA, 1998.
- [12] Malay K. Pakhira, "A Modified k-means Algorithm to Avoid Empty", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 220-226, 2009.
- [13] Meila, M. and Heckerman, D. "An experimental comparison of model-based clustering methods", Machine Learning, kluwer Academic publishers, Vol. 42, pp. 9-29, 2001.
- [14] Miha Grcar, Marko Grobelnik and Dunja Mladenic, "Using Text Mining and Link Analysis for Software Mining", Lecture Notes in Computer Science, Vol. 4944, pp. 1-12, 2008.
- [15] Murtagh, F. "A Survey of Recent Advances in Hierarchical Clustering Algorithms Which Use Cluster Centers", Comput. J, Vol. 26, pp. 354-359, 1984
- [16] Pallav Roxy and Durga Toshniwal, "Clustering Unstructured Text Documents Using Fading Function", International Journal of Information and Mathematical Sciences, Vol. 5, No. 3, pp. 149-156, 2009
- [17] Shehroz S. Khan and Amir Ahmad, "Cluster Center Initialization Algorithm for K-means Clustering", Pattern Recognition Letters, Vol. 25, No. 11, pp. 1293-1302, 2004.
- [18] Shin-Jye Lee and Xiao-Jun Zeng, "A three-part input-output clustering-based approach to fuzzy system identification", 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 55-60, 2010.
- [19] Ward Jr, J.H. "Hierarchical grouping to optimize an objective function", J. Am. Stat. Association, Vol. 58, pp. 236-244, 1963.