International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Performance Analysis in Text Clustering Technique

[1]Ms. D. Devikanniga, [2]Dr. M. Thangamani

[1]Assistant Professor (Selection Grade), Sri Ramakrishna Engineering College, Coimbatore, nsuresh2@gmail.com
[2]Assistant Professor, Kongu Engineering College, Perundurai-638052, manithangamani2@gmail.com

Abstract: *Document clustering is becoming more and more important with the abundance of text documents available through World Wide Web and corporate document management systems. 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. This survey includes the information about data mining clustering technique for unstructured data.*

Keywords: Data Mining, Text mining, Document clustering and Clustering techniques.

1. Introduction

Text clustering is one of the vital processes in text mining. Liping [1] 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.

2. Clustering Techniques

(i) Global K-means

Likas *et al.* [2] 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 [3] stipulated iterative clustering techniques to calculate initial cluster centers for K-means. This process is feasible for clustering techniques for continuous data.

(ii) CLIQUE Algorithm

Agrawal *et al.* [4] 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.

A clustering algorithm called CLIQUE fulfills all the above requirements. CLIQUE finds dense clusters in subspaces of maximum dimensionality. It produces cluster explanations in the form of Disjunctive Normal Function (DNF) expressions that are reduced for ease of comprehension. The approach generates matching results without considering the data in which input records are presented. It does not presume any particular mathematical form for data distribution. From the experimental result, it was observed that CLIQUE algorithm efficiently identified accurate clusters in large high dimensional datasets.

(iii) Enhanced K-means

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 [5] 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.

(iv) Heterogeneous Uncertainty Clustering Feature (H-UCF)
Uncertainty heterogeneous data streams [6] 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.* [7] 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.

(v) Hierarchical Particle Swarm Optimization (HPSO) clustering
Alam *et al.* [8] 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.

(vi) Hybrid clustering approach (HCA)
Shin-Jye Lee *et al.*[9] 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

(vii) Incremental clustering
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.* [10] 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. Then clustering ensemble was employed to combine or partition clusters to gain optimal clustering. Ultimately, the proposed approach was applied in yellow-small data set, diagnosis data set and zoo data set and results revealed the effectiveness of the approach.

(viii) Partitional clustering approach
Clustering approaches have been extensively used in the area of pattern discovery [11] from Web Usage Data (WUD). In e—commerce applications, clustering is used for the purpose of creating marketing policies, product offerings etc. Raju *et al.* [12] presented a novel partition based technique for dynamically grouping web users, based on their web access patterns, using Adaptive Resonance Theory Neural Network (ART1NN) clustering algorithm. The results show that ART1NN clustering technique outperforms K-means and Self Organizing Map (SOM) clustering algorithms in terms of intra-cluster and inter-cluster distances.

3. Text clustering

(i) Data grabber
Crescenzi *et al* .[13] 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.

(ii) Link-Analysis and Text-mining toolbox (LATINO)
Miha Grcar *et al.* [14] 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.* [15]. LATINO, an open source purpose data mining platform, offers text mining, link analysis, machine learning, etc.

(iii) Similarity and model based approaches
Similarity-based approach and model-based approaches [16] are the two major categories of clustering approaches and these have been described by Pallav Roxy and Durga Toshniwal [17]. 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.

Self-organizing map [18], mixture of Gaussians [19], spherical K-mean [20], bi-secting K-means [21], mixture of multinomial [22] are some of the new techniques available to improve clustering performance. K-means, an unsupervised learning technique, is good at solving clustering issues as well as minimizing objective function. It defines K-centroids for each cluster. After

positioning, the centroids located away from each other, where each point equals to a given set, should be taken and linked next to the centroid. Then, it is allowed to recompute K-centroids. Next, by means of a loop, the location of K-centroids is altered until no more alterations are made.

(iv) Fuzzy clustering for text

Jiabin Deng *et al.* [23] came up with an enhanced fuzzy clustering text clustering by relying on the Fuzzy C-Means (FCM) and edit distance approach. It uses feature estimation and minimizes dimensionality of a high-dimensional text vector. The researchers, in order to sustain the stability of FCM output, added aspects like high-power sample point set, field of radius and weight. Since FCM has constraints such as boundary value attribution [24] they recommended the edit distance approach. Consequently, the outputs showed that it was more precise and constant than FCM in regard to text clustering.

Odukoya *et al.* [25] organized an enhanced data clustering approach for mining web documents which formulates, simulates and assesses the web documents with the intention of conserving their conceptual relationships. The enhanced data clustering approach was formulated with the concept of K-means algorithm. The experimental result showed that this clustering approach attains more accuracy (89.3%) than the other existing approach (88.9%). The entropy was constant for both approaches with a value of 0.2485 at k = 3. This also reduces with the increase in the number of clusters until the number of clusters reaches eight where it increases to some extent. The altered rand index values change from 0 to 1 for both clustering approaches.

The other existing approach attains a value of 53% when compared with this approach which attains an altered rand index value of 63.7%, when the number of clusters was five. Additionally, the response time got reduced from 0.0451 seconds to 0.0439 seconds when the number of clusters was three. This confirms that the response time of the data clustering approach has been reduced by 2.7% when compared with the traditional K-means data clustering. This study revealed that the proposed data clustering approach could be utilized by developers of web search engines for well-organized web search result clustering.

(v) Semantic concepts

In a way, a majority of the available clustering algorithms in the Chinese text clustering suffer from the drawbacks of data scalability and the interpretability of results. Liu Jinling and Zhou Hong [26] presented a well-organized Chinese text clustering technique depending on the semantic concepts. This technique significantly reduced the number of required data to be processed and enhanced the capability of the clustering approaches. The experimental output showed that this clustering approach attained an acceptable clustering outcome and better implementation efficiency.

(vi) Seeds Affinity Propagation

Renchu Guan *et al.* [27] presented a novel semisupervised Seeds Affinity Propagation (SAP) approach based on an Affinity Propagation (AP) algorithm. The two most important contributions in this technique are:

- Structural information of texts are obtained by a novel similarity metric.
- The semisupervised clustering process is enhanced by seed construction approach.

From the experimental results, it is revealed that the proposed similarity metric is more efficient in text clustering and the proposed semi supervised approach attains better clustering outcomes and faster convergence (using only 76 % iterations of the original AP). The entire SAP algorithm attains a higher F-measure and lower entropy; improves considerable clustering execution time (20 times faster) with respect to K-means, and offers better robustness when compared with other existing approaches.

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. 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 [28]. 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 [29]. 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. Even though several approaches for clustering exist, most of these feature vectors have the same principal problems without really approaching the matters of subjectivity and explanation. So the researcher intends to investigate various aggregation levels. From that perspective, the text documents will be used to obtain clustering results.

Big data technologies defines a new generation of technologies and architectures, designed solely to economically extract useful information's from very large volumes of a wide variety of data, by permitting high velocity capture, discovery, and analysis. Big data is the data that exceeds the processing capacity of conventional database systems. Big data is very big, moves very fast, or doesn't fit into traditional database architectures. There are different definitions of big data as it is more often used as an all-encompassing term for everything from actual data sets to big data technology and big data analytics.

There are mainly 3 types of big data sets- structured, semi structured and unstructured. In structured data, we can group the data to form a relational schema and represent it using rows and columns within a standard database. Based on an organization's parameters and operational needs, structured data responds to simple queries and provides usable information due to its configuration and consistency. Semi structured data does not conform to an explicit and fixed schema.

Document clustering is the use of cluster analysis of textual documents. It has many applications like organizing large document collection, finding similar documents, recommendation system, duplicate content detection, search optimization. Document clustering has been considered for use in a number of different areas of text mining and information retrieval. There are many search engines that are using for information retrieval, but the main challenge in front of the search engine is to present relevant results of the user. Even though there are many knowledge discovery tools to filter, order, classify or cluster their search results exists, still user make extra effects to find the required document. In order to provide solution, combining the entire web mining based data mining techniques. Now a day's huge data is producing by many social networking websites, e-commerce websites, and many organizations. Analyzing this huge data is tedious task for any organization. To analyses the huge data, database management techniques may not be sufficient, so the big data came into existence.

References

[1] Liping Jing, "Survey of Text Clustering", Department of Mathematics, The University of Hong Kong, HongKong, China, ISBN: 7695-1754-4/02, 2005

[2] Likas, A., Vlassis, N. and Verbeek, J.J. "The Global k-means Clustering algorithm", Pattern Recognition , Vol. 36, No. 2, pp. 451-461, 2003.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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] 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.

[9] 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.

[10] 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.

[11] Nagy, G. "State of the art in pattern recognition", Proceedings IEEE, Vol. 56, pp. 836-862, 1968.

[12] Raju, G.T. and Sudhamani, M.V. "A novel approach for extraction of cluster patterns from Web Usage Data and its performance analysis", International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), pp. 718-723, 2011.

[13] Crescenzi valter, Giansalvatore Mecca, Paolo Merialdo and Paolo Missier, "An Automatic Data Grabber for Large Web Sites", VLDB , pp. 1321-1324, 2004.

[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] 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.

[16] Meila, M. and Heckerman, D. "An experimental comparison of model-based clustering methods", Machine Learning, kluwer Academic publishers, Vol. 42, pp. 9-29, 2001.

[17] 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.

[18] Kohonen, T., Kaski, S., Lagus, K., Salojrvi, J., Honkela, J., Paatero, V. and Saarela, A. "Self Organization of a Massive Document Collection", IEEE Transactions Neural Networks, Vol. 11, pp. 574-585, 2000.

[19] Tantrum, J., Murua, A. and Stuetzle, W. "Hierarchical model-based clustering of large datasets through fractionation and refractionation", Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 183-190, 2002.

[20] Dhillon Inderjit, S. and Modha Dharmendra, S. "A data clustering algorithm on distributed memory multiprocessors", In Large-Scale Parallel Data Mining, pp. 245-260, 2000.

[21] Steinbach, M., Karypis, G. and Kumar,V. "A Comparison of Document Clustering Techniques", KDD Workshop on Text Mining, pp. 109-110, 2000.

[22] Vaithyanathan, S. and Dom, B. "Model-based Hierarchical Clustering", Proc. 16th Conf. Uncertainty in Artificial Intelligence, pp. 599-608, 2000.

[23] Jiabin Deng, JuanLi Hu, Hehua Chi and Juebo Wu, "An Improved Fuzzy Clustering Method for Text Mining", Second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC), Vol. 1, pp. 65-69, 2010.

[24] Dave, R.N. "Generalized fuzzy C-shells Clustering and Detection of Circular and Elliptic Boundaries", Pattern Recognition, Vol. 25, pp. 713-722, 1992.

[25] Odukoya, O.H., Aderounmu, G.A. and Adagunodo, E.R. "An Improved Data Clustering Algorithm for Mining Web Documents", International Conference on Computational Intelligence and Software Engineering (CiSE), pp. 1-8, 2010.

[26] Liu Jinling and Zhou Hong, "Clustering Efficient Method on Mass Chinese Text Based on Semantic Concept", International Forum on Information Technology and Applications (IFITA), Vol. 2, pp. 151-155, 2010.

[27] Renchu Guan, Xiaohu Shi, Marchese, M., Chen Yang and Yanchun Liang, "Text Clustering with Seeds Affinity Propagation", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 4, pp. 627 - 637, 2011.

[28] 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.

[29] Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U. "When is 'Nearest Neighbor' Meaningful", In Proc. of ICDT-1999, Jerusalem, Israel, pp. 217-235, 1999.

Author Profile

Mrs. D.Devikanniga is currently working as Assistant Professor in the Department of Information Technology at Sri Ramakrishna Engineering College, Coimbatore. She has a Bachelor's degree in Information Technology (2003), a Master's degree in Advanced Computing (2005). She has above 9 years of teaching experience and guided 16 UG projects. Her research and teaching interests include Data Mining, Cloud computing, Big Data, Opinion Mining and Sentimental Analysis, Data Structures, Design and Analysis of Algorithms. She is a life time member in ISTE and IAENG. She published papers in National and International Conferences and Journals.

Dr. M. Thangamani is nearly 20 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published 50 articles in International journals and presented over 67 papers in national and international conferences in above field. She has delivered more than 35 Guest Lectures in reputed engineering colleges on various topics. She has organized many self supporting and sponsored national conference and Workshop in the field of data mining, big data and cloud computing. She is on the editorial board and reviewing committee of leading research journals, and on the program committee of top international data mining and soft computing conferences in various countries. She also seasonal reviewer in IEEE Transaction on Fuzzy System, international journal of advances in Fuzzy System and Applied mathematics and information journals. She has organizing chair and keynote speaker in international conferences in India and abroad. She is currently working as Assistant Professor in Kongu Engineering College.