International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Review of Modern Document Clustering Algorithm

Priti B. Kudal[1], Prof. Manisha Naoghare[2]

[1]Student, Master of Engineering, Department of Computer Engineering Sir Visvesvaraya Institute of Technology, Chincholi, Sinner

[2]Assistant Professor, Department of Computer Engineering Sir Visvesvaraya Institute of Technology, Chincholi, Sinner

Abstract: ***It is important to emphasize that getting from a collection of documents to a clustering of the collection, is not merely a single operation, but is more a process in multiple stages. These stages include more traditional information retrieval operations such as crawling, indexing, weighting, filtering etc. Some of these other processes are central to the quality and performance of most clustering algorithms, and it is thus necessary to consider these stages together with a given clustering algorithm to harness its true potential. We will give a brief overview of the clustering process, before we begin our literature study and analysis.***

Keywords: Clustering, Data Mining

1. Introduction

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document cluster in scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document clustering.

Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification. No super-vision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. Clustering is sometimes erroneously referred to as automatic classification; however, this is inaccurate, since the clusters found are not known prior to processing whereas in case of classification the classes are pre-defined. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so called training set, i.e. a set of data correctly labeled by hand, and then replicates the learnt behavior on unlabeled data.

2. Literature Survey

K-means is the most important flat clustering algorithm. The objective function of K-means is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C:

The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. A measure of how well the centroids represent the members of their clusters is the Residual Sum of Squares (RSS), the squared distance of each vector from its centroid summed over all vectors. K-means can start with selecting as initial clusters centers K randomly chosen objects, namely the seeds. It then moves the cluster centers around in space in order to minimize RSS. This is done iteratively by repeating two steps until a stopping criterion is met

1) Reassigning objects to the cluster with closest centroid
2) Recomputing each centroid based on the current members of its cluster.

We can use one of the following termination conditions as stopping criterion:

- A fixed number of iterations I has been completed.
- Centroids μ do not change between iterations.
- Terminate when RSS falls below a pre-estabilished threshold.

2.1 Expectation Maximization

The EM algorithm fall within a subcategory of the flat clustering algorithms, called Model-based clustering. The model-based clustering assumes that data were generated by a model and then tries to recover the original model from the data. This model then defines clusters and the cluster membership of data. The EM algorithm is a generalization of K-Means algorithm in which the set of K centroids as the model that generate the data. It alternates between an expectation step, corresponding to reassignment, and a maximization step, corresponding to recomputation of the parameters of the model.

2.2 Hierarchical Clustering

Hierarchical clustering approaches attempt to create a hierarchical decomposition of the given document collection thus achieving a hierarchical structure. Hierarchical methods are usually classified into Agglomerative and Divisive methods depending on how the hierarchy is constructed.

Agglomerative methods start with an initial clustering of the term space, where all documents are considered representing

a separate cluster. The closest clusters using a given inter-cluster similarity measures a r e then merged continuously until only 1 cluster or a predefined number of clusters remain. Simple Agglomerative Clustering Algorithm:

1) Compute the similarity between all pairs of clusters i.e. calculates a similarity matrix whose ij entry gives the similarity between the i and j clusters.
2) Merge the most similar (closest) two clusters.
3) Update the similarity matrix to reflect the pair wise similarity between the new cluster and the original clusters.
4) Repeat steps 2 and 3 until only a single cluster remains.

Divisive clustering algorithms start with a single cluster containing all documents. It then continuously divides clusters until all documents are contained in their own cluster or a predefined number of clusters are found. Agglomerative algorithms are usually classified according to the inter-cluster similarity measure they use. The most popular of these are single-link, complete-link and group average. In the single link method, the distance between clusters is the minimum distance between any pair of elements drawn from these clusters (one from each), in the complete link it is the maximum distance and in the average link it is correspondingly an average distance.

2.3 Other Algorithms

There are only a few studies reporting the use of clustering algorithms in the Computer Forensics field. Essentially, most of the studies describe the use of classic algorithms for clustering data—e.g., Expectation-

Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM). These algorithms have well-known properties and are widely used in practice. For instance, K-means and FCM can be seen as particular cases of EM [8]. Algorithms like SOM [24], in their turn, generally have inductive biases similar to K-means, but are usually less computationally efficient. In [6], SOM-based algorithms were used for clustering files with the aim of making the decision-making process performed by the examiners more efficient. The files were clustered by taking into account their creation dates/times and their extensions. This kind of algorithm has also been used in [20] in order to cluster the results from keyword searches. The underlying assumption is that the clustered results can increase the information retrieval efficiency, because it would not be necessary to review all the documents found by the user anymore. An integrated environment for mining e-mails for forensic analysis, using classification and clustering algorithms, was presented in [21]. In a related application domain, e-mails are grouped by using lexical, syntactic, structural, and domain-specific features [10]. Three clustering algorithms (K-means, Bisecting K-means and EM) were used. The problem of clustering e-mails for forensic analysis was also addressed in [23], where a Kernel-based variant of K-means was applied. The obtained results were analyzed subjectively, and the authors concluded that they are interesting and useful from an investigation perspective. More recently [14], a FCM-based method for mining association rules from forensic data was described.

The literature on Computer Forensics only reports the use of algorithms that assume that the number of clusters is known and fixed apriori by the user. Aimed at relaxing this assumption, which is often unrealistic in practical applications, a common approach in other domains involves estimating the number of clusters from data. Essentially, one induces different data partitions (with different numbers of clusters) and then assesses them with a relative validity index in order to estimate the best value for the number of clusters [7], [3], [19]. This work makes use of such methods, thus potentially facilitating the work of the expert examiner—who in practice would hardly know the number of clusters apriori.

3. Conclusion

In this paper we investigated many existing algorithms. We conclude that it is hardly possible to get a general algorithm, which can work the best in clustering all types of datasets. Thus we plan to implement two algorithms which can work well. Thus, these algorithms can be very effective in applications like a search engine for a particular field. Finally we would conclude that though many algorithms have been proposed for clustering but still an open problem.

References

[1] L. N. Fred and A. K. Jain, Combining multiple clusterings using evidence accumulation,‖ IEEE Trans.Pattern Anal. Mach. Intell., vol. 27, no. 6, pp. 835– 850, Jun. 2005.
[2] Strehl and J. Ghosh, ―Cluster ensembles: A knowledge reuse framework for combining multiple partitions,‖ J. Mach. Learning Res., vol. 3, pp. 583–617, 2002.
[3] A.K.JainandR.C.Dubes, Algorithms for Clustering Data.Englewood Cliffs, NJ: Prentice-Hall, 1988.
[4] Aggarwal, C. C. Charu, and C. X. Zhai, Eds., Chapter4:ASurvey of Text Clustering Algorithms,‖ in Mining Text Data.NewYork: Springer, 2012.
[5] Mirkin, Clustering for Data Mining: A Data Recovery Approach. London, U.K.: Chapman & Hall, 2005.
[6] B.K.L. Fei,J.H.P.Eloff, H.S.Venter, and M.S. Oliver, Exploring forensic data with self-organizing maps,‖ in
[7] Proc. IFIP Int. Conf. Digital Forensics, 2005, pp. 113–123. 2008 [7] B.S.Everitt,S.Landau, and M.Leese, Cluster Analysis. London, U.K.: Arnold, 2001.
[8] M. Bishop, Pattern Recognition and Machine Learning.New York: Springer- Verlag, 2006.
[9] R. Hruschka, R. J. G. B. Campello, and L. N. de Castro, ‖Evol ving clusters in gene-expression data,‖Inf. Sci., vol. 176, pp. 1898–1927, 2006.
[10] Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, -Mining write prints from anonymous e-mails for forensic investigation, Digital Investigation, Elsevier, vol. 7, no. 1–2, pp. 56–64, 2010.
[11] Salton and C. Buckley, ―Term weighting approaches in automatic text retrieval,‖ Inf. Process. Manage, vol. 24, no. 5, pp. 513–523, 1988.

[12] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz, ―The expanding digital universe: A forecast of worldwide information growth through 2010,‖ Inf. Data, vol. 1, pp. 1–21, 2007.
[13] K. Kishida, - High-speed rough clustering for very large document collections J.Amer. Soc. Inf. Sci., vol. 61, pp. 1092– 1104, 2010, doi: 10.1002/ asi.2131.
[14] K. Stoffel, P. Cotofrei, and D. Han, ―Fuzzy methods for forensic data analysis,‖ in Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition, 2010, pp. 23–28.
[15] L. F. Nassif and E. R. Hruschka, ―Document clustering for forensic computing: An approach for improving computer inspection,‖ in Proc. Tenth Int. Conf. Machine Learning and Applications (ICMLA),2011, vol. 1, pp. 265–268, IEEE Press.
[16] L. Hubert and P. Arabie, Comparing partitions,‖ J. Classification, vol. 2, pp. 193–218, 1985.
[17] L. Kaufman and P. Rousseeuw, Finding Groups in Gata: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.
[18] L. Liu, J. Kang, J. Yu, and Z. Wang, ―A comparative study on unsupervised feature selection methods for text clustering,‖ in Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering, 2005, pp. 597–601.
[19] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, ―Relative clustering validity criteria: A comparative overview,‖ Statist. Anal. Data Mining, vol. 3, pp. 209–235, 2010.
[20] N. L. Beebe and J. G. Clark, ―Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results, Digital Investigation, Elsevier, vol. 4, no. 1, pp. 49–54, 2007.
[21] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, ―Towards an integrated e-mail forensic analysis framework,‖ Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137, 2009.
[22] R.XuandD.C.Wunsch,II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009.
[23] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, ―Text clustering for digital forensics analysis,‖ Computat. Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009.
[24] S. Haykin, Neural Networks: A Comprehensive Foundation. Englewood Cliffs, NJ: Prentice-Hall, 1998.
[25] S. Nassar, J. Sander, and C. Cheng, ―Incremental and effective data summarization for dynamic hierarchical clustering,‖ in Proc. 2004 ACM SIGMOD Int. Conf. Management of Data (SIGMOD ’04), 2004, pp. 467–478.
[26] V. Levenshtein, ―Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady, vol. 10, pp. 707–710, 1966.
[27] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial, ―Effcient algorithms for exact hierarchical clustering of huge datasets: Tackling the entire protein space, Bioinformatics, vol. 24, no. 13, pp. i41–i49, 2008.
[28] Y. Zhao and G. Karypis, ―Evaluation of hierarchical clustering algorithms for document datasets,‖ in Proc. CIKM, 2002, pp. 515–524.
[29] Y. Zhao, G. Karypis, and U. M. Fayyad,―Hierarchical clustering algorithms for document datasets,‖ Data Min. Knowl. Discov., vol. 10, no. 2, pp. 141–168, 2005.

Author Profile

Priti Kudal received the B.E. degree in Computer Engineering from Sir Visvesvaraya Institute of Technology in 2008. She is currently pursuing Masters Degree in Computer.

M. M. Naoghar is working as Assistant Professor in Sir Visvesvaraya Institute of Technology.