Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.332 – 335

SURVEY ARTICLE

A Survey on Text Based Clustering

S.Chidambaranathan

Head, Department of MCA, St. Xavier's College (Autonomous), Tirunelveli, India

scharan2009@rediffmail.com

Abstract— ***Clustering is the main technique for data analysis and it deals with the organisation of a set of objects in a multidimensional space into cohesive groups called clusters. Every cluster contains closely related objects and has very dissimilar objects in other clusters. Cluster analysis aims at discovering the objects with same behaviour in a collection. Thus, if an object satisfies a rule, the all objects that are similar to that object will satisfy the same rule. With this functionality the hidden similarity, relationship and other concepts can be predicted with respect to the cluster of objects. In this work, we present a survey on text based mining.***

Keywords— ***Clustering, text based mining***

I. INTRODUCTION

We are in the internet era which is overloaded with much data. The growth of data is directly proportional to time. Thus, the web search experience is not feasible to many average users.

The dynamic and heterogeneous nature of the web in association with data overload, leads to the problem of poor data retrieval.

It is essential to increase the degree of ease in data retrieval, such that the browsing experience of a user will be enhanced.

In order to make the data retrieval easier for the users, search engines, meta-search engines and web directories are developed. The user who hunts for information submits a query to the search engine and the search engine matches the query and the key terms and the result is provided to the user.

The problem arises when the query submitted by the user is irrelevant, which paves way to the presentation of several irrelevant web pages. It is better to have a technique that effectively organizes and presents the most relevant pages to the user. This objective is attained by the technique 'document clustering'. Because of its importance, several algorithms are being proposed.

Clustering is the main technique for data analysis and it deals with the organisation of a set of objects in a multidimensional space into cohesive groups called clusters [1]. Every cluster contains closely related objects and has very dissimilar objects in other clusters. Cluster analysis aims at discovering the objects with same behaviour in a collection.

Thus, if an object satisfies a rule, the all objects that are similar to that object will satisfy the same rule. With this functionality the hidden similarity, relationship and other concepts can be predicted with respect to the cluster of objects.

To cluster a document, the type of attributes such as the words or phrases upon which the clustering algorithm lies has to be figured out.

II. WEB DOCUMENT CLUSTERING

A clustering algorithm is based on the attribute with which a cluster is to be formed. They are categorized into text based clustering, link based clustering and hybrid clustering. In the text based clustering, the clustering relies on the content of the document. Link based clustering is based on the linked pages in a collection and finally the hybrid clustering takes both text and link based clustering into account.

2.1 TEXT BASED CLUSTERING

Text based clustering is determined by the content of the document. To be fine, when two documents are found to possess many common words, then obviously those are similar documents. The text based approaches are further classified into partitional, hierarchical, graph-based, neural-network based and probabilistic approaches. Most of the existing text based clustering algorithms are stringent, meaning that a document belongs to a cluster or not.

2.1.1 Partitional Clustering

A partitional or a non-hierarchical clustering partitions a collection of documents into a predefined number of disjoint clusters. These partitional clustering can further be classified into iterative or single pass methods. The iterative methods can also be called as reallocation methods. Single pass methods are usually employed in the beginning of the reallocation methods, for generating the first partitioned data.

A feature vector matrix is used by the partitional clustering algorithms and the clusters are produced by the optimization of a criterion function. Some of the criterion functions are maximizing the sum of the average pairwise cosine similarities between the documents assigned to a cluster and minimizing the cosine similarity of each cluster centroid to the centroid of the whole collection and so on. The clustering solution is affected by the criterion function and so much concentration has to be rendered in the selection of criterion function.

The best example for partitional clustering algorithm is k-means, which relies on the centroid. The centroid is the central point of the cluster and this can be the good representative of the cluster. Initially, this algorithm selects k-cluster centroids. After this, the cosine distance between each of the document in the cluster and the centroid is calculated. Then, the document is assigned to the cluster with the nearest centroid. After assigning all the documents to a cluster, the new cluster centroids are calculated and this happens repetitively until a criterion is met.

The main advantages of this algorithm are its simplicity and their low computational complexity. The drawback of this algorithm is its parameter dependency such as the target number of clusters, initial cluster centroid selection and the document processing order.

2.1.2 Hierarchical Clustering

This algorithm produces the sequence of nested partitions. An nxn similarity matrix is used to store the similarity between each pair of documents. At every step, the algorithm either merges two clusters or splits a cluster into two. The clustering results are displayed with a tree like structure namely, dendrogram. This contains a cluster as a root which contains all the documents of the collection and the clusters at the bottom contains at least a document. The step-by-step procedure of hierarchical clustering is given below.

1. Assign each document to a single cluster.
2. Compute the similarity between all pairs of clusters and store the result in a similarity matrix, in which the ij-th entry stores the similarity between the i-th and j-th cluster.
3. Merge the two most similar (closest) clusters.
4. Update the similarity matrix with the similarity between the new cluster and the original clusters.
5. Repeat steps 3 and 4 until only one cluster remains or until a threshold is reached.

Some of the popular approaches of this clustering are single link, complete link, group average and ward's method.

Single Link

The similarity between a pair of clusters is calculated on the basis of similarity between the two most similar documents, each in a cluster. This method generates loosely bound clusters with little internal cohesion. This algorithm exploits several useful mathematical properties and lesser computational complexity. Some of the single link's algorithms are van Rijsbergen's algorithm, SLINK, Minimal Spanning Tree and Voorhees's algorithm.

Complete Link:

The similarity between a pair of clusters is taken to be the similarity between the least similar documents, one of which is in each cluster. This is much stricter than that of the single link method. Here, the clusters are small and tightly bound. Some of the algorithms that are on the basis of complete link are CLINK algorithm, which is a variation of the SLINK algorithm, and the algorithm proposed by Voorhees.

Group Average:

This method produces clusters such that each document in a cluster has greater average similarity with the other documents in the cluster than with the documents in any other cluster. All the documents in a cluster contribute in the calculation of the pairwise similarity and, thus, this method is a mid-point between the above two methods.

Ward's Method

In this method, the cluster pair to be merged is the one whose merger minimizes the increase in the total within-group error sum of squares based on the distance between the cluster centroids. This method tends to result in spherical, tightly bound clusters and is less sensitive to outliers. Ward's method can be implemented using the reciprocal-nearest neighbour (RNN) algorithm, which was modified for document clustering by Handouchi and Willett.

Centroid/Median Method:

Each cluster formed is represented by the group centroid/median. At each stage of the clustering, the pair of clusters with the most similar mean centroid/median is merged. The difference between the centroid and the median is that the second is not weighted proportionally to the size of the cluster.

HAC approaches produce high quality clusters but have very high computational requirements and are greedy. This means that a pair of clusters that is chosen for agglomeration, at each time, is the one, which is considered the best at that time, without regard to future consequences. Also, if a merge that has taken place is not appropriate, there is no backtracking to correct the mistake.

2.1.3 Graph based Clustering

In this case, the documents need to be clustered can be viewed as a set of nodes and the edges between the nodes represent the relationship between them. The edges bare a weight, which denotes the strength of the relationship. Graph based algorithms rely on graph partitioning, that is, they identify the clusters by cutting edges from the graph such that the edge-cut, i.e. the sum of the weights of the edges that are cut, is minimized. The basic idea is that the weights of the edges in the same cluster will be greater than the weights of the edges across clusters. Hence, the resulting cluster will contain highly related documents. Different graph based algorithms differ by the graph production and the graph partitioning algorithm being used. Chameleon's graph representation of the document set is based on the k-nearest neighbour graph approach.

2.1.4 Neural Network based Clustering

The Kohonen's Self-Organizing feature Maps (SOM) (Kohonen, 1995) is a widely used unsupervised neural network model. It consists of two layers: the input layer with n input nodes, which correspond to the n documents, and an output layer with k output nodes, which correspond to k decision regions (i.e. clusters). The input units receive the input data and propagate them onto the output units. Each of the k output units is assigned a weight vector. During each learning step, a document from the collection is associated with the output node, which has the most similar weight vector. The weight vector of that 'winner' node is then adapted in such a way that it will become even more similar to the vector that represents that document, i.e. the weight vector of the output node 'moves closer' to the feature vector of the document. This process runs iteratively until there are no more changes in the weight vectors of the output nodes. The output of the algorithm is the arrangement of the input documents in a 2- dimensional space in such a way that the similarity between the input documents is mirrored in terms of topographic distance between the k decision regions.

2.1.5 Fuzzy Clustering

All the aforementioned approaches produce clusters in which, each document is assigned to one and only one cluster. Fuzzy clustering approaches, on the other hand, are non-exclusive, in the sense that each document can belong to more than one cluster. Fuzzy algorithms usually try to find the best clustering by optimising a certain criterion function. The fact that a document can belong to more than one clusters is described by a membership function. The membership function computes a membership vector for each document, in which the i-th element indicates the degree of membership of the document in the i-th cluster. The most widely used fuzzy clustering algorithm is Fuzzy c-means, a variation of the partitional k-means algorithm. In fuzzy c-means each cluster is represented by a *cluster prototype* (the center of the cluster) and the membership degree of a document to each cluster depends on the distance between the document and each cluster prototype.

The closest the document is to a cluster prototype, the greater is the membership degree of the document in the cluster. Another fuzzy approach, that tries to overcome the fact that fuzzy c-means doesn't take into account the distribution of the document vectors in each cluster, is the Fuzzy Clustering and Fuzzy Merging algorithm (FCFM) (Looney, 1999). The FCFM uses Gaussian weighted feature vectors to represent the cluster prototypes. If a document vector is equally close to two prototypes, then it belongs more to the widely distributed cluster than to the narrowly distributed cluster.

2.1.6 Probability Clustering

Another way of dealing with uncertainty is the usage of probabilistic clustering algorithms. These algorithms use statistical models to calculate the similarity between the data instead of some predefined measures. The basic idea is the assignment of probabilities for the membership of a document in a cluster. Each document may belong to more than one cluster according to the probability of belonging to each cluster. Probabilistic clustering approaches are based on finite mixture modeling (Everitt and Hand, 1981).

They assume that the data can be partitioned into clusters that are characterized by a probability distribution function (p.d.f.). The p.d.f. of a cluster gives the probability of observing a document with particular weight values on its feature vector in that cluster. Since the membership of a document in each cluster is not

known a priori, the data are characterised by a distribution, which is the mixture of all the cluster distributions. Two widely used probabilistic algorithms are Expectation Maximization (EM) and AutoClass (Cheeseman and Stutz, 1996). The output of the probabilistic algorithms is the set of distribution function parameter values and the probability of membership of each document to each cluster.

III. CONCLUSIONS

Text based clustering is determined by the content of the document. To be fine, when two documents are found to possess many common words, then obviously those are similar documents. The text based approaches are further classified into partitional, hierarchical, graph-based, neural-network based and probabilistic approaches. Thus, this work provides a detailed survey of text based mining.

REFERENCES

[1] Bezdek, J.C., Ehrlich, R., Full, W. 1984. FCM: Fuzzy C-Means Algorithm. Computers and Geosciences.

[2] Boley, D., Gini, M., Gross, R., Han, E.H., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., Moore, J. 1999. Partitioning-based clustering for web document categorization. Decision Support Systems, 27(3):329-341.

[3] Botafogo, R.A., Shneiderman, B. 1991. Identifying aggregates in hypertext structures. Proc. 3rd ACM Conference on Hypertext pp.63-74.

[4] Botafogo, R.A. 1993. Cluster analysis for hypertext systems. Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, pp.116-125

[5] Cheeseman, P., Stutz, J. 1996. Bayesian Classification (AutoClass): Theory and Results. Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, pp. 153-180.

[6] Croft, W. B. 1993. Retrieval strategies for hypertext. Information Processing and Management, 29:313-324

[7] Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W. 1992. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, pp.318-329.

[8] Defays, D. 1977. An efficient algorithm for the complete link method. The Computer Journal, 20:364-366.

[9] Dhillon, I.S. 2001. Co-clustering documents and words using Bipartite Spectral Graph Partitioning. UT CS Technical Report # TR 2001-05

[10] Ding, Y. 2001. *IR and AI: The role of ontology.* Proc. 4th International Conference of Asian Digital Libraries, Bangalore, India

[11] El-Hamdouchi, A., Willett, P. 1986. Hierarchic document clustering using Ward's method. Proceedings of the Ninth International Conference on Research and Development in Information Retrieval. ACM, Washington, pp.149-156

[12] El-Hamdouchi, A., Willett, P. 1989. Comparison of hierarchic agglomerative clustering methods for document retrieval. The Computer Journal 32

[13] Frei, H. P., Stieger, D. 1995. The Use of Semantic Links in Hypertext Information Retrieval. Information Processing and Management,pp 1-13.

[14] Han, E.H., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., Moore, J. 1997. WebACE: a web agent for document categorization and exploration. Technical Report TR-97-049, Department of Computer Science, University of Minnesota, Minneapolis (http://www.users.cs.umn.edu/~karypis/publications/ir.html)

Author Profile

S.Chidambaranathan, received his post graduate degree in Mathematics from Madurai Kamaraj University, Madurai. He also earned post graduate degree in Computer Application and M.Phil in computer Science from Manonmaniam Sundaranar University, Tirunelveli. Presently he is working as HoD in the Department of MCA, St. Xavier's College (Autonomous), Palayamkottai, Tamil Nadu. He is an author for many books including "PHP for beginners", "XML-An Practical approach" and "Everything HTML". He has published many research papers in National, International journals and conference proceedings.

He can be contacted at E-mail: scharan2009@rediffmail.com

