METADATA BASED CLUSTERING MODEL FOR DATA MINING

[1]T. NADANA RAVISHANKAR, [2]R. SHRIRAM

[1] Research Scholar., Department of Computer Science and Engineering, B.S.A University, Chennai -48.

[2]Professor., Department of Computer Science and Engineering, B.S.A University, Chennai -48.

E-mail: [1]nadanaravishankar@gmail.com, [2]Shriram@bsauniv.ac.in

ABSTRACT

As the information flooding onto worldwide web is growing fast, it is very difficult to find the hidden knowledge from huge data stored. In traditional clustering techniques, the task for clustering algorithm is to find the relevant data, describing relationships among the elements between two input datasets. During unsupervised learning process, metadata may play a significant role in the context of data retrieval. In this work, we proposed a methodology of metadata based clustering model (MBCM) for large datasets. The proposed model is validated with few standard datasets like IEEE, ACM and Cluto etc. Experimental results show that it is possible to achieve better cluster quality without significant overhead in terms of execution time. Finally, the performance of our proposed model is evaluated using F-measure and the performance of our method is compared with existing clustering models.

Keywords: *Metadata, clustering, data mining, dataset, K medoids.*

1. INTRODUCTION

Nowadays, information overloading is a major problem with a growing number of users and information sources. Search engines like Google, Yahoo etc., try to produce solutions for customer need by using clustering algorithms. Clustering algorithm has been utilized to group similar objects together into a cluster for better searching and interpreting of data. In large data sets, a major problem is that there is more information hidden in the quantity that can be envisioned or tacit. It is very difficult for a user to have an entire copy of the large data sets on their system. It is also very difficult to find the relevant information from complex data set that you need. Traditional clustering algorithms retrieve only what user specify in the query. An effective method for information retrieval against large data sets is the iterative process of querying and refining.

Metadata is precise information that makes it easier to retrieve or manage an information resource. Metadata is often called data about data. For example, traditional library arrangement is a form of metadata. The metadata has potential to enhance open data users and publishers [1].There are three main types of metadata available: 1) *Descriptive metadata* describes a resource for purposes such as discovery and identification, 2) *Structural metadata* indicates how compound objects are put together, for example, how pages are ordered to form chapters, and 3) *Administrative metadata* provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. We use descriptive metadata for our implementation. The major reason for creating descriptive metadata is to facilitate discovery of relevant information by browsing or query. Examples of descriptive metadata are identifier, title, publisher, country, source, type, format, language, sector, subjects, keywords, relative information system, validity date (from – to), relevant resources and linked data sets.

Our goal is to find the relevant data to user query from a large database. Metadata can also useful to organize electronic resources, facilitate interoperability and legacy resource integration, and provide digital identification, and support archiving and safeguarding. Thus the success of the clustering approaches is dependent largely on the expressive nature of the metadata. This research work trusts methodology which can be used to extract metadata during a clustering process. In the context of this experimental evaluation and application study we have obtained promising results.

The rest of this paper is organized as follows. Section 2 analyses the salient prior work on metadata based clustering models. Section 2 and 3 discusses the problem statement and proposed methodology of MBCM. Section 4 describes the implementation and the results. Section 5 concludes the work with a summary of our findings and a discussion of issues that could be effectively reconnoitered in future work.

2. RELATED WORK

Data mining is a new and on-going research domain, which needs efficient clustering methods to find information in a large data set is an important problem. In the recent past [2], they used RDF to model some elements of metadata to describe the resources without making assumptions about a particular application domain. In [3, 4], they proposed hybrid mining models for classification. Apriori is a well-known algorithm which is used extensively in market-basket analysis and data mining. The algorithm is used for learning association rules from transactional databases and is based on simple counting procedures. In hybrid model, Apriori is further improved by C4.5 decision tree and K-means clustering algorithms, respectively.

Metadata is also used to embed available contextual knowledge into the co-clustering process [5]. They proposed some metadata strategies to leverage the available metadata in discovering contextually-relevant co-clusters, without significant overheads in terms of information theoretical co-cluster quality or execution cost.

The recent text mining research shows that effective usage and update of discovered patterns is still an open research issue [6, 7 and 8]. To improve the effectiveness of using and updating discovered text patterns for finding relevant and interesting information, this study proposes effective models by utilizing metadata. Using Support Vector Machines, metadata can be automatically extracted from a set of documents [9]. Automatically extracting header features is applicable to the research idea, because the header information is part of the metadata. Metadata can also be applied on image clustering [10], in that they proposed a robust method for those missing metadata of photo images that appear in search results on the web.

In [11], they extended Scatter/Gather to support user interactions and adjustments to the clusters. They developed an effective prototype, but have limitations if the data set is absurdly large. The clusters are viewed in a tree structure and grouped hierarchically, allowing the user to drill-down into subgroups. This capability to zoom into a cluster, or increase granularity is quite useful. In [12], they made a prototype defined as Interactive Visual Clustering and make an attempt to cluster the data to address a user's goals. The interface makes an update to operate on user interactions with the clusters in two dimensions. Each of the nodes can be dragged by the user to a different "more appropriate" cluster. Once a user has moved two nodes, the clustering automatically readjusts. The data is clustered based on its attributes.

3. PROBLEM STATEMENT

Using metadata, we can manage a huge volume of data. They can guarantee both the cluster quality as well as the reliability of the data retrieved. In this paper, we present a methodology which can be used to extract metadata during a retrieval process. Our goal is to determine how metadata can best be used to improve clustering large datasets. In large database, inclusion of metadata provides well separation of data objects from large dataset in order to facilitate thematic clustering of resources [13]. However, clustering algorithms are difficult to evaluate. Manual evaluations of cluster quality are time consuming and usually not well suited for comparing across many different algorithms or settings [14]. Most of the clustering engines are developed only for document clustering. However, a good clustering engine will effectively have to retrieve and rank the search result not only for text and should be effective for dataset as well. The primary objective of this work is to demonstrate a methodology for Metadata based Clustering which improves the relevancy of the retrieved content from the large datasets.

4. PROPOSED METHOD

We can distinguish two main phases in our proposed system: a) the metadata estimation phase, where the dataset is processed and organized, and; b) the clustering algorithm phase, where the dataset's to be labelled are retrieved and ranked in order to decide their output clusters. The overall methodology of the proposed approach is shown in Figure 1.

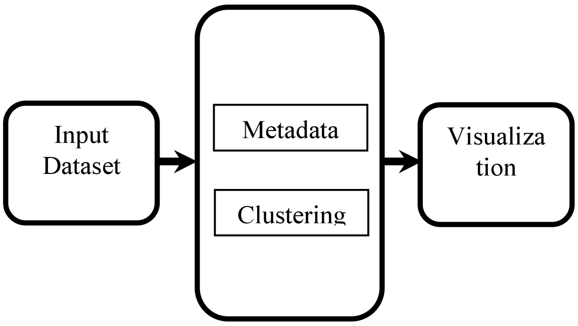

Figure 1: The Proposed Method of MBCM

4.1 Metadata Estimation Phase

In this section, we describe how metadata can be extracted from large dataset. Though there are various structures of metadata, the main goal is to provide access to resources given by providing all information concern to them. Using metadata, we can improve accessibility of data, discovery of data, interpretation of data, linking data, avoid unnecessary duplication of open data and business objects and business processes. However they may also have some drawbacks like time-consuming, high costs of adding and maintaining metadata etc.

Our approach estimates metadata of a data item d_i in an input dataset D = {d_1, d_2,…., d_n} using Euclidean distance metric. We estimate two kinds of metadata: local l_i, and global g_i, that the data d_i has.

The details of our method for metadata estimation are as follows; to estimate the metadata of a dataset, we use "datastage 8.1" proprietary software from IBM. For example, a simple dataset (food.xls) is given to the software as input. The result that contains the metadata estimated for the data denoted by D'. Table 1 shows the simple 'food.xls' dataset with 5 attributes such as energy, protein, calcium, fat and iron. We can estimate the metadata for a food item or an attribute. Table 2 shows the metadata estimated for the data item 'Chicken broiled'. We call this metadata as local that contains the details only about one food item. To estimate the metadata, for example, for attribute protein, we have to select the protein as input to estimate metadata. Table 3 shows the metadata estimated for protein. We can also estimate the metadata for multiple items at a single run using the same software.

Table 1: Example Dataset For Metadata Estimation

Food	Energy	Protein	Fat	Calcium	Iron
Beef	340	20	28	9	2.6
Hamburger	245	21	17	9	2.7
Beef Roast	420	15	39	7	2
Beef Steak	375	19	32	9	2.6
Beef	180	22	10	17	3.7
Chicken	115	20	3	8	1.4
Chicken	170	25	7	12	1.5
Beef Heart	160	26	5	14	5.9
Lamb Leg	265	20	20	9	2.6
Lamb	300	18	25	9	2.3
Smoked	340	20	28	9	2.5
Pork Roast	340	19	29	9	2.5
Beef	205	18	14	7	2.5
Veal	185	23	9	9	2.7
Bluefish	135	22	4	25	0.6
Clams	70	11	1	82	6
Clams	45	7	1	74	5.4
Crabmeat	90	14	2	38	0.8
Mackerel	200	19	13	5	1
Mackerel	155	16	9	157	1.8
Perch	195	16	11	14	1.3
Salmon	120	17	5	159	0.7
Sardines	180	22	9	367	2.5

Table 2: Example For Local Metadata

Food	Energy	Protein	Fat	Calcium	Iron
Chicken Broiled	115	20	3	8	1.4
	Normal	Normal	Low	Low	Low

. *Table 3: Example for Estimated Metadata for Protein*

	Low	Normal	Abnormal
Protein	Clams Raw	Beef Braised	Hamburger
	Clams	Beef Roast	Beef
	Crabmeat	Beef Steak	Chicken
		Chicken	Beef Heart
		Lamb Leg	Veal Cutlet
		Lamb	Bluefish
		Smoked Ham	Sardines
		Pork Roast	Tuna
		Pork	Shrimp
		Beef Tongue	
		Haddock Fried	
		Mackerel	
		Mackerel	
		Perch Fried	

4.2 The Clustering Algorithm

We have applied K-medoids algorithm to produce the results of the manipulation technology, is a variation of the K-means algorithm. The K-means algorithm is a distinguished technique for performing clustering on objects in large datasets [15, 16]. K-means can be applied to the clustering of objects but there is a reason for not doing so. In the K-means algorithm, each cluster is represented by the center of the cluster and distances to the centroids are calculated for each object at every iteration. Since the original database may have tens of thousands of records, such designs will be slow. K-means is efficient for large data sets but sensitive to outliers and the clustering results for different similarity measures are more or less the same regardless of the size of the dataset [17]. However Euclidean distance metric works well for large database in order to find distance between set of data objects [18].

In K-medoids algorithm, each cluster is represented by one of the objects in the cluster itself and the algorithm is not sensitive to outliers. The steps involved in K-medoids algorithm is described below:

Step 1: Initial *k* representatives are chosen randomly.
Step 2: Cluster each point based on the closest Center.
Step 3: Replace each centre by the medoid of points in its cluster.
Step 4: Stop when within-cluster variation doesn't change.

For example, Table 4 shows the simple dataset for clustering. Initially, we have started with 2 clusters. We formed the following clusters using Euclidean distance metric,

Cluster 1 = $\{O_1, O_2, O_3, O_4\}$
Cluster 2 = $\{O_5, O_6, O_7, O_8, O_9, O_{10}\}$

Table 4: Simple dataset for clustering

Data Objects	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
A1	2	3	3	4	6	6	7	7	8	7
A2	6	4	8	7	2	4	3	4	5	6

The algorithm minimizes the sum of the dissimilarities between each object (O_i) and its corresponding reference point (P). The idea of the algorithm is to minimize the sum of absolute error (A.E) for all objects in the data set that is given in Eq. (1).

$$A.E = \sum_{i-1}^{n} \sum_{P \in Ci} |P - Oi| \qquad (1)$$

Figure 2 shows the clustered scatter diagram for the applied dataset. In this, we choose to divide the objects into two clusters. Red coloured object denotes the medoid of the cluster.

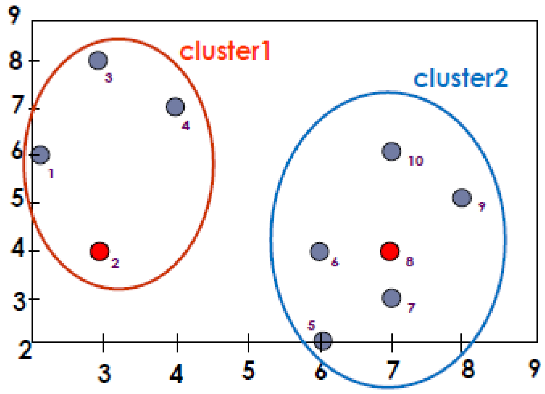

Figure 2: Scatter Diagram for K – Medoids Clustering

5. RESULTS AND ANALYSIS

The entire algorithm is implemented in Weka3.7 software. The proposed work implemented and compared with existing K-means clustering method on the assumption that metadata is estimated for all the data's in a dataset. To ensure diversity of the research work, we have chosen commonly used datasets from ACM, IEEE and Cluto, for evaluating clusters as the experiment datasets. For each estimated datasets, we obtained the cluster result with the help of K-medoids algorithm. The number of clusters is set equal to the number of manually assigned labels. Finally, the quality of clustering results is estimated using commonly used measures such as precision (P), recall (R) and F - measure.

Precision is a measure of how many of the objects retrieved by the query were judged relevant and recall is how many of the relevant documents we retrieved versus how many there were in the database. The precision (P) and recall (R) of a cluster ‘i’ with respect to a class ‘c’ are defined in the following Eq. (3) and Eq. (4):

$$P = \frac{O_{ci}}{O_i} \qquad (2)$$

$$R = \frac{O_{ci}}{O_i} \qquad (3)$$

F – measure (F_c) is a measure that combines P and R. From Equations (2) and (3), the F – measure of class ‘c’ is calculated based on Eqn. (4).

$$F_c = \frac{2\,P.R}{P+R} \qquad (4)$$

Table 5: Precision of Existing Vs MBCM

Dataset	Existing method	MBCM
ACM	0.698	0.821
CLUTO	0.765	0.854
IEEE	0.724	0.865

Table 6: Recall of Existing Vs MBCM

Dataset	Existing method	MBCM
ACM	0.616	0.697
CLUTO	0.712	0.795
IEEE	0.651	0.714

Table 7: Comparison of F – Measure of Existing Vs MBCM

Dataset	Existing method	MBCM
ACM	0.654	0.753
CLUTO	0.738	0.823
IEEE	0.686	0.782

From Table 5 and 6, it is clear that the precision and recall are better than existing system in the proposed MBCM. The comparison of F-measure of existing methods and proposed MBCM are shown in the Table 7 and its graphical representation is shown in figure 3. Using metadata, we can improve better cluster quality without significant time overheads.

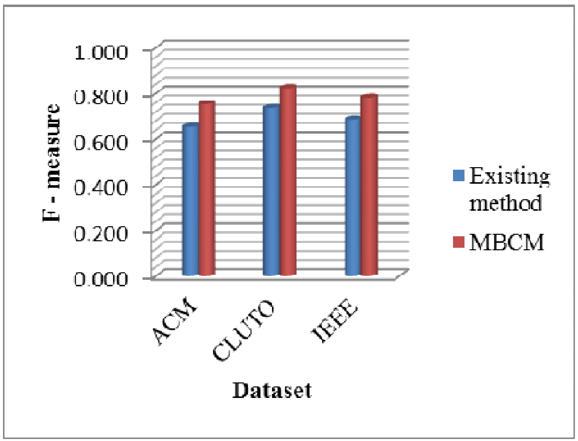

Figure 3: F – Measure Comparison

It is observed that our metadata based clustering algorithm performs better and the performance improvement over simple clustering methods is about 10% from Table 7. By inclusion of metadata we can achieve better cluster quality and relevant relevance of the data retrieved. However, extracting metadata for large dataset is time consuming.

6. CONCLUSION

In this work, we have presented approach towards data mining using metadata based clustering method to provide users with better search results. The main contribution of this work is to improve the cluster quality and reliability of the retrieved data without any overheads. We estimate the metadata in the search results and empirically evaluated our approach and compared the performance of our approach to the base line methods. The experimental results have shown that our metadata based clustering model is effective process than existing models. In our future work, Clustering based on metadata presented in this paper can be applied for documents and real-world datasets. Another future work is to develop a prototype for concept mining based clustering scheme under the direction of metadata and study the performance.

REFRENCES:

[1] Anneke Zuiderwijk, Keith Jeffery, and Marijn Janssen 2012. "The potential of metadata for linked open data and its value for users and publishers". Journal of Democracy, Vol.4, No.2, ISSN 2075-9517, pp. 222-244.

[2] Abdourahamane Balde and Marie-Aude Aufaure 2005. "How can metadata contribute to add semantic information to clusters", The Electronic Journal of Symbolic Data Analysis - Vol.3, No.1, ISSN 1723-5081, 2005, pp. 32-44.

[3] Brown, S. and B. Forouraghi 2009. "Concept Classification using a hybrid data mining model". 21st International Conference on Tools with Artificial Intelligence, Nov. 2-4, IEEE Xplore Press, Newark, NJ. pp. 375-378.

[4] Rahman, S.M.M., M.R.A. Kotwal and Y. Xinghuo 2010. "Mining classification rules via an apriori approach". Proceedings of the 13th International Conference on Computer and Information Technology, Dec. 23-25, IEEE Xplore Press, Dhaka, pp. 388-393.

[5] Claudio Schifanella Maria Luisa Sapino and K. Selçuk Candan 2012. "On context-aware co-clustering with metadata support". Journal of Intelligent Information Systems, Volume 38, Issue-1, pp. 209-239.

[6] Zhong, N.Y. Li and S.T. Wu 2010. "Effective pattern discovery for text mining". IEEE Transactions on Knowledge and Data Engineering, 24: pp. 30-44.

[7] Koteeswaran. S, J. Janet and E. Kannan 2012. "Significant Term List Based Metadata Conceptual Mining Model for Effective Text Clustering". Journal of Computer Science Vol. 8 No. 10, pp. 1660-1666.

[8] Kam-Fai Wong, Nam-Kiu Chan, and Kam-Lai Wong 2003. "Improving Document Clustering by Utilizing Meta-Data". In the proceedings of the sixth international workshop on information retrieval with Asian languages-AsianIR'03, Vol.11, pp. 109-115.

[9] Hui Han, C. Lee Giles, Eren Manavoglu, Hongyuan Zha, Zhenyue Zhang, and Edward 2003. Fox. "Automatic document metadata extraction using support vector machines". In Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, JCDL '03, pp. 37–48, Washington, DC, USA.

[10] Masaharu Hirotha, Naoki Fukuta, Shohei Yokoyama, and Hiroshi Ishikawa 2012. "A robust clustering method for missing metadata in image search results". Journal of Information Processing, Vol.20, No.3, pp. 537-547.

[11] Omar Alonso and Justin Talbot 2008. "Structuring collections with scatter/gather extensions". In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, New York, NY, USA, pp. 697–698.

[12] Marie Desjardins, James MacGlashan, and Julia Ferraioli 2007. "Interactive visual clustering". In Proceedings of the 12th international conference on Intelligent user interfaces, IUI '07, New York, USA, pp. 361–364.

[13] J. Lacasta, J. Nogueras-Iso, P.R. Muro-Medrano, and F.J. Zarazaga-Soria 2007. "Thematic clustering of geographic resource metadata collections". In Web and Wireless Geographical Information Systems, Lecture Notes in Computer Science (LNCS), Vol. 4857, pp. 30-43.

[14] T. Haveliwala, A. Gionis, D. Klein, and P. Indyk 2002. "Evaluating strategies for similarity search on the web". In the Proceedings of the 11th International conference on World Wide Web WWW '02, pp. 432-444.

[15] Sandhia Valsala, Jissy Ann, and George Priyanka Parvathy 2011. "A Study of Clustering and Classification Algorithms Used in Data mining". International Journal of Computer Science and Network Security (IJCSNS), VOL.11 No.10, pp. 167-174.

[16] Osama Abu Abbas 2008. "Comparisons between data clustering algorithms", The International Arab Journal 0f Information Technology, Vol.5, No.3, pp. 320-325.

[17] Anna Huang 2008. "Similarity Measures for Text Document Clustering". In the proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC), pp. 49-56.

[18] Nadana Ravishankar and Shriram 2012. "Metrics for Information Retrieval: A Case Study". In Proceedings of the International Conference on Software Engineering and Mobile Application Modelling and Development, ICSEMA 2012, pp. 40-48, India.

[19] Miroslav Cepek and Pavel Kordiki 2012 Dataset Clustering According to Metadata Extracted from Dataset CTU Technical Report.