Journal of Computer Science 8 (10): 1660-1666, 2012
ISSN 1549-3636
© 2012 Science Publications

Significant Term List Based Metadata Conceptual Mining Model for Effective Text Clustering

[1]Koteeswaran, S., [2]J. Janet and [1]E. Kannan
[1]Department of CSE,
Vel Tech Dr. RR and Dr. SR Technical University, Chennai, India
[2]Department of CSE and IT, Dr. MGR University, Chennai, India

Abstract: As the engineering world are growing fast, the usage of data for the day to day activity of the engineering industry also growing rapidly. In order to handle and to find the hidden knowledge from huge data storage, data mining is very helpful right now. Text mining, network mining, multimedia mining, trend analysis are few applications of data mining. In text mining, there are variety of methods are proposed by many researchers, even though high precision, better recall are still is a critical issues. In this study, text mining is focused and conceptual mining model is applied for improved clustering in the text mining. The proposed work is termed as Meta data Conceptual Mining Model (MCMM), is validated with few world leading technical digital library data sets such as IEEE, ACM and Scopus. The performance derived as precision, recall are described in terms of Entropy, F-Measure which are calculated and compared with existing term based model and concept based mining model.

Key words: Meta data conceptual mining model, clustering, text mining, data mining

INTRODUCTION

Data mining is an iterative knowledge model to discover hidden knowledge through either automatic or manual methods. Data mining is the most useful field of study, in which new, valuable and nontrivial information in large volumes of data are handled by innovative and efficient methodologies.

The major tasks (Kantardzic, 2003) in the data mining are, Classification-discovery of a predictive learning function that classifies a data item into one of several predefined classes; Regression-discovery of a predictive learning function, which maps a data item to a real-value prediction variable; Clustering-a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data; Summarization-an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data; Dependency Modelling-finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set; Change and Deviation Detection-discovering the most significant changes in the data set.

In the above, the focus of recent research in the data mining is further reduced as clustering, prediction and the classification. The Prediction is the process which predicts unknown or future values of interest by using some variables or fields in the data set and the prediction produces the model of the system described. Classification is the process which is used for finding patterns by describing the data that can be interpreted and the classification produces new, nontrivial information based on the available data set. In order to execute these processes in the data mining requires clustering and outlier analysis for reducing as well as identifying useful dataset.

Cluster analysis is a methodology for classifying given samples into a number of defined groups using a pre-defined measure of association. Therefore, the samples in one group are similar and the samples belonging to different groups are dissimilar. Simply says, when a set of samples and a measure of similarity (or dissimilarity) between two samples are given as input to the clustering model, which return number of groups (clusters) that form a partition, or a structure of partitions, of the data set.

Mathematical model of clustering and literature survey: Consider that an ordered pair (X, *s*), or (X, *d*) are input samples, where X is a set of descriptions of samples and *s* and *d* are measures for similarity or dissimilarity between the samples, respectively. The output of the clustering system is a partition $A = \{G_1, G_2, \ldots, G_N\}$ where G_k, $k = 1, \ldots, N$ is a crisp subset of X such that Eq 2:

Corresponding Author: Koteeswaran. S., Department of CSE, Vel Tech Dr. RR and Dr.SR Technical University, Chennai

$$G_1 \cup G_2 \cup \ldots, \cup G_N = X \quad (1)$$

And

$$G_1 \cap G_2 \cap \ldots, \cap G_N = \emptyset \quad (2)$$

The G_1, G_2 … G_n are the clusters.

The clustering is processed using Quantitative features and Qualitative features. The Quantitative features can be subdivided as (1) continuous values (e.g., real numbers where $P_j \subseteq R$), (2) discrete values (e.g., binary numbers $P_j=\{0, 1\}$, or integers $P_j \subseteq Z$) and 3) interval values (e.g., $P_j = \{x_{ij} \leq 20, 20 < x_{ij} < 40, x_{ij} \geq 40\}$. The Qualitative features can be subdivided as (1) nominal or unorderedvalues (e.g., color is "blue" or "red") and (2) ordinalvalues (e.g., military rank with values "general", "colonel",).

The word "similarity" in clustering means that the value of s (x, x') is large when x and x' are two similar samples; the value of s (x, x') is small when x and x' are not similar. Very often a measure of dissimilarity is used instead of a similarity measure. A dissimilarity measure is denoted by d (x, x'), $\forall x$, $x' \in X$. Dissimilarity is frequently called a distance. A distance d (x, x') is small when x and x' are similar; if x and x' are not similar d (x, x') is large.

It is obvious that when p = 1, then d coincides with L_1 distance and when p = 2, d is identical with the Euclidean metric. For example, for 4-dimensional vectors x_1 = {l, 0, 1, 0} and x_2 = {2, 1, - 3, - 1} these distance measures are $d_1 = 1+1+4+1 = 7$, $d_2 = (1+1+16+1)^{1/2} = 4.36$ and $d_3 = (1 + 1+64+1)^{1/3} = 4.06$.

Text mining is a new and on-going research domain, which needs efficient clustering methods. In initial stages of data mining research, various classifiers using association rules are applied for knowledge discovery. Most of the classifiers uses positive rues as similarity measures. Kundu *et al*. (2008) proposes negative rules for associative classifier. The generation of negative associations from datasets has been attacked from different perspectives by various authors and this has proved to be a very computationally expensive task. The authors proposes the classifier, which termed as "Associative Classifier with Negative rules"(ACN) is not only time-efficient but also achieves significantly better accuracy than four other state-of-the-art classification methods by experimenting on benchmark UCI datasets.

The comparison shown by Mazid *et al*. (2009) gives the detailed study of Association ruled based mining model. In which the Rule based mining (which may be performed through either supervised learning or unsupervised learning techniques) are compared with recent research proposals using predefined test sets. In terms of accuracy and computational complexity, the author concluded Apriori is a better choice for rule based mining task.

Later on 2009, hybrid mining model are proposed for classification, for ex, concept classification proposed by Brown and Forouraghi, (2009) and Rahman *et al*. (2010). As already concluded that, 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.

El-far *et al*. (2011) proposed k-means classifier for data mining which applied for Three-dimensional data models to visualize realistic objects. This study is proposes k-means for application such as CAD/CAO, medical simulations, games, virtual reality. There are two major approaches for drawing or building 3d objects, (1) the search in the database can be done via requests that are either 3D objects, (2) via some 2D views of the 3D object. This study contributes an extract characteristic views of 3D models using Data Mining algorithms which comprises Apriori, Charm, Close+ and Extraction of association rules. The work tested using a database that contains 120 numbers of 3D models selected from the Princeton Shape Benchmark, for 342numbers of 2D views.

The recent text mining research shows that effective usage and update of discovered patterns is still an open research issue (Zhong *et al*, 2012a). To improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information, this study proposes effective pattern methods. For detailed survey of text mining, for clustering (Koteeswaran *et al*, 2012), a survey of evolutionary algorithm by Barros *et al* (2012) and survey of twenty of years of mixure of experts by Yuksel 2012 are recommended.

The concept based mining model proposed by (Shehata and Kamel, 2010), used concept based analysis for text clustering. The concept on the sentence, documents and corpus levels rather than a single term analysis on the document are the objective of this study. The Conceptual Term Frequency (CTF) in sentences, Term Frequency (TF) are calculated and based on these calculation, the text are classified as particular nature.

This was further modified by Cai *et al*. (2012), in which the authors used Nonnegative Matrix Factorization for text categorization. NMF can only be performed in the originalfeature space of the data points and it gives acceptable result than existing system.

Pattern taxonomy for text classification (Zhong *et al*, 2012b) proposes closed sequential patterns, which used the well-known Apriori property in order to reduce the searching space. The advancement of DBSCAN, named TSCAN (Chen and Chen, 2012) defined an event as a significant theme development that continues fora period of time. In general, all these events are temporally disjoint and which may be taken together form the message of the topic. Moreover, events in different themes may beassociated because of theirtemporal proximity and context similarity. The authors proposes a model to identifies the themes and the events from the given documents and associatedevents.

The recent development of conceptual text mining includes string mining which concentrates low memory usage (Dhaliwal *et al*., 2012), Text deduction methodology (Chenghua *et al*, 2012) which proposes a novel probabilistic modeling framework called Joint Sentiment-Topic (JST) model based on Latent Dirichl*et al*location (LDA) are recommended implementation of recent research.

Proposed work: The k-means algorithm uses number of terms appeared in the documents, based on these calculations, the documents sorting the list of terms which appeared most frequently in the documents. The terms are filtered and analyzed by a technical person for categorizing the documents. So that it needs technical person for clustering for accurate manipulations.

In the Term Based Method proposes by Li *et al*. (2000), information retrieval provided many using rough set method or support vector machine based filtering model. The advantages of term based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

In order to avoid technical person interpretation and manipulation, as it involves more costly job, the concept based mining model proposes concept analysis. In the concept analysis, the ctf, tf are calculated and based on these calculations higher ctf and tf are sorted. These most frequent ctf and tf terms are verified with the technical terms which prepared in the preprocessing stage. Therefore, it needs any clerical level staff member to classify the documents.

In the proposed work, the Meta data Conceptual Mining Model (MCMM), used for effective text classification. The proposed MCMM executes in two s of manipulation, which are training phase and testing phase as shown in Fig. 1.

The proposed MCMM are explained in the following.

Training phase: In the preprocessing stage, the di-grams (such as in, as, it) and tri-grams (such as are, for, ing) terms are removed from the documents.

Significant Term List (STL) is a list of keywords which prepared by a technical person based on their domain of study. STL are prepared one each for each field of study, i.e., each clustering groups. The STL which has basic terminology will be updated each time, the text is clustered. And the STL has unique, primary key terms which appeared in only one STL and it will not re-appear in another.

In the conceptual analysis stage, the terms which appeared in each STL are searched in the given training documents.

The ctfvalues of the documents are shown in the Eq. 1, ctf = number of frequent terms/total number of terms in the documents-(1)

In the classification stage, the highest values of ctf which appeared in any one field of STL is identified and clustered as the name of STL. This process continues for each training documents and each additional relevant terms identified in the training phase is added in the concern STL.

Testing phase: Similar to training phase, in the preprocessing stage, the di-grams (such as in, as, it) and tri-grams (such as are, for, ing) terms are removed from the documents.

In the conceptual analysis stage, the ctf values of each term which appeared in every STL are calculated from the given document.

In the classification stage, the highest values of ctf which appeared in any one field of STL is identified and clustered as the name of STL.

MCMM Algorithm: The algorithm of proposed work which explained in the above section is given in the following sub-section:

A. Training Phase

Step 1: Apply preprocessing (remove di-grams and tri-grams)

Step 2: Prepare Significant Term List (STL) for each field of study

Step 3: Check the metadata stored in each STL is unique and primary data

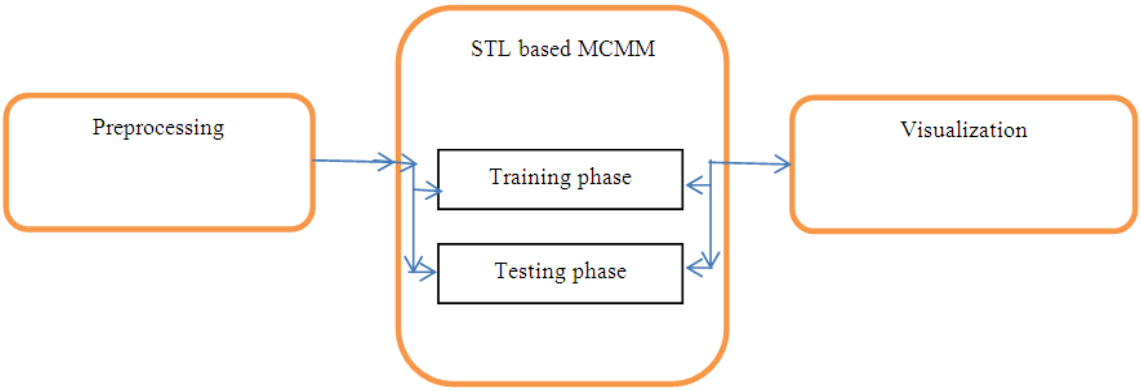

Fig. 1: Design of proposed MCMM

Step 4: Read training documents until all training documents are read otherwise goto step 9.
Step 5: Calculate the number of matching terms in the given documents which matching the STL are 'm' and calculate the total number of sentences in the given documents are 'n'
Step 6: Apply Concept Analysis model for finding ctf, ctf = m/n
Step 7: Sort the ctf in decreasing order and check the terms which has higher ctfare available in the STL, if available goto step 8 otherwise gotostep 9.
Step 8: Update these new terms to concern STL and goto step 3
Step 9: End the training process

B. Testing Phase

Step 1: Apply preprocessing (remove di-grams and tri-grams)
Step 2: Collect Significant Term List (STL) for each field of study
Step 3: Check the metadata stored in each STL is unique and primary data
Step 4: Apply the input test document
Step 5: Read each term in every STL and Calculate the number of matching terms in the given test document which matching the STL are 'm' and Calculate the total number of sentences in the given test document is 'n'
Step 6: Calculate ctf, ctf = m/n
Step 7: Sort the ctf in decreasing order,
Step 8: Check the terms which has higher ctf
Step 9: Check this highest ctf term is available in the given STL, if available goto step 10 otherwise gotostep11.
Step 10: Classify the given test document as the field of matching STL
Step 11: Identify next higher ctf term until ctf become zero and goto Step 9

RESULTS

The manipulation methodologies implemented in k-means algorithm, concept based model and proposed methods are shown in the Table 1. The base technique shows the methodology used for clustering, the classification shows the mode of operation used for clustering and the stage shows the implementation paradigm of each methodology, the last metric is given answer for the preprocessing is implemented in each methodology.

The proposed work implemented and compared with Term based method and concept based method. The result of the implementation are recorded and shown in the Table 1 and 2. The inputs are collected from the world leading technical study consortium such as IEEE, ACM and Scopus. The IEEE is a collection of technical data base which available online through IEEE Explore. The IEEE Explore is a digital library and search engine which contains high quality of technical articles from international conference proceedings and transactions. The ACM is also a high quality digital library which contains technical articles from varies ACM Transactions. The Scopus has world largest collection of technical articles which contains almost all leading technical and management journals like IEEE, ACM, Elsevier, Willey, Oxford, Springer and Taylor-Francis.

The accuracy and performance of text mining is measured using two measures, namely F-measure and entropy. The F-measure is the metric used for measuring performance of the clustering technique, which is calculated based on following Eq. 3-6.

The F-measure is calculation which combines the precision and recall function from the information retrieval procedure.

The precision P of a cluster 'j' with respect to a class 'i' are defined in the following Eq. 3:

Table 1: Comparison of manipulation models of existing Vs proposed methods

Description	Term based model	Concept based model	Proposed MCMM
Base technique	Frequent itemset	Concept term frequency	Concept term frequency, Significant term list
Performance updating	Not possible	Not possible	Possible. By proper training, the performance may be achieved as higher as possible.
Stage	One Stage	One Stage	Two Stage (Training and Testing)
Preprocessing Support	Yes	Yes	Yes
No of Search	n×n	n×n	T×n ($\forall$ T < n)
	ex: 150×150	ex: 150×150	ex: 10 x 150

Table 2: Comparison of F-Measure of existing Vs proposed methods

Field of Study	Data Set	Term based model	Concept based model	Proposed MCMM
Electrical	IEEE	0.697	0.741	0.823
	ACM	0.767	0.812	0.876
	Scopus	0.724	0.807	0.859
Electronics	IEEE	0.688	0.731	0.812
	ACM	0.757	0.801	0.865
	Scopus	0.715	0.797	0.848
Civil	IEEE	0.756	0.804	0.892
	ACM	0.832	0.881	0.950
	Scopus	0.785	0.875	0.932
Computer	IEEE	0.746	0.793	0.881
	ACM	0.821	0.869	0.938
	Scopus	0.775	0.864	0.919
Mechanical	IEEE	0.736	0.783	0.869
	ACM	0.810	0.858	0.925
	Scopus	0.765	0.853	0.907

Table 3: Comparison of Entropy of existing methods Vs proposed methods

Field of Study	Data Set	Term based model	Concept based model	Proposed MCMM
Electrical	IEEE	0.329	0.214	0.143
	ACM	0.317	0.178	0.132
	Scopus	0.412	0.380	0.297
Electronics	IEEE	0.325	0.211	0.141
	ACM	0.313	0.176	0.130
	Scopus	0.407	0.375	0.293
Civil	IEEE	0.357	0.232	0.155
	ACM	0.344	0.193	0.143
	Scopus	0.447	0.412	0.322
Computer	IEEE	0.352	0.229	0.153
	ACM	0.339	0.191	0.141
	Scopus	0.441	0.407	0.318
Mechanical	IEEE	0.348	0.226	0.151
	ACM	0.335	0.188	0.139
	Scopus	0.435	0.401	0.314

$$P = \text{precion}(i, j) = \frac{Mij}{Mj} \quad (3)$$

The recall function R of a cluster 'j' with respect to a class 'i' am defined in the following Eq. 4:

$$R = \text{recall}(i, j) = \frac{Mij}{Mj} \quad (4)$$

where, M_{ij} is the number of members of a class 'i' in a cluster 'j', M_j is the number of members of class 'i'.

From the Eq. 3 and 4, the F-Measure of a class 'i' is defined in the following Eq. 5:

$$F(i) = \frac{2 * P * R}{P + R} \quad (5)$$

The overall F-measure is calculated based on the following Eq. 6:

$$F = \frac{\sum_i \left(|i| \times F(i)\right)}{\sum_i |i|} \quad (6)$$

The comparison of F-measure of various existing methods and proposed MCMM are shown in the Table 2.

The one more metric of performance calculation for text mining is Entropy, which explained in the following Eq. 7 and 8.

The Entropy is a measure of quality for untested clusters, which also defined as quality of clusters at one level of a hierarchical clustering. Entropy measures the homogeneous of a cluster and the higher the homogeneous of a cluster replies the lowest entropy of the cluster. Suppose, the cluster has perfect homogeneity, the entropy of the concern cluster becomes zero.

The Entropy of class 'i' is defined in the following Eq. 7:

$$E_j = -\sum_i^j p_{ij} \times \log(p_{ij}) \quad (7)$$

The overall Entropy of cluster is defined in the following Eq. 8:

$$E = \sum_{j=1}^{n} \left(\frac{M_j}{M} \times E_j \right) \quad (8)$$

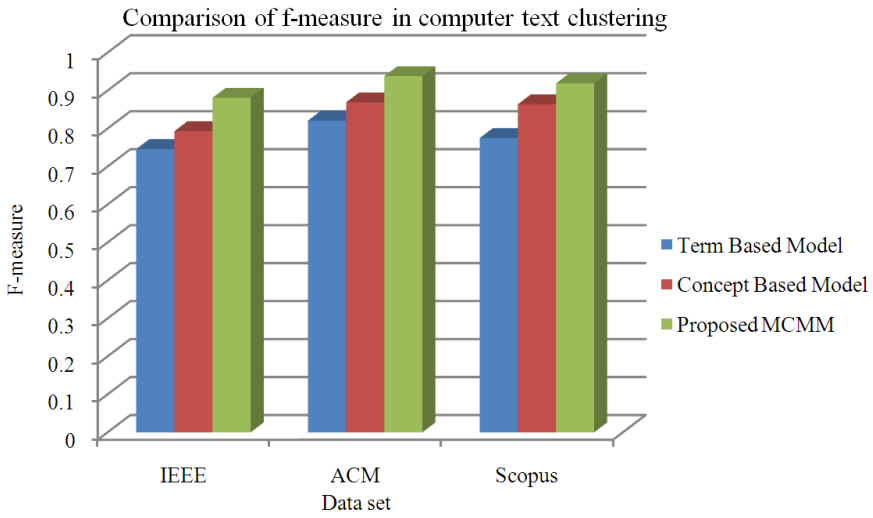

Fig. 2: Comparison of F-Measure of existing Vs proposed methods

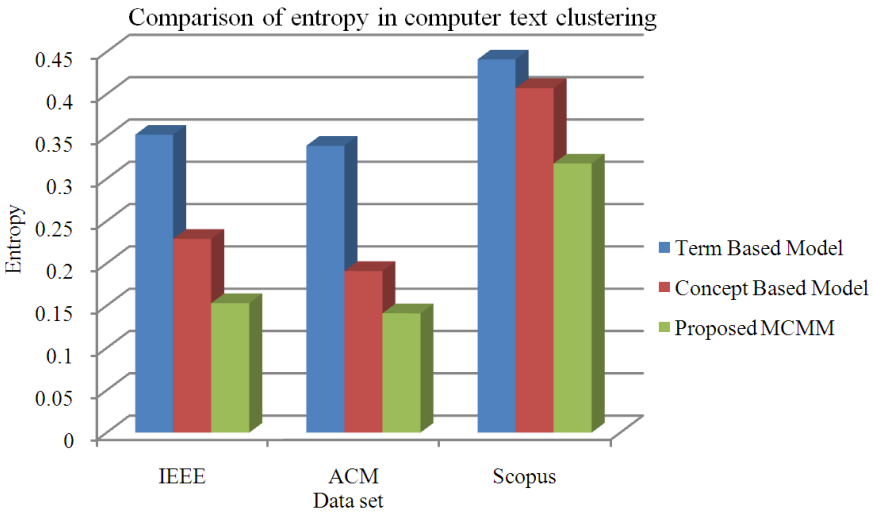

Fig. 3: Comparison of Entropy of existing Vs proposed methods

The Entropy of the existing and proposed methods is displayed in the Table 3.

The graphical representation of performance result which shown in the Table 2 and 3 are shown in the Fig. 2 and 3.

CONCLUSION

The Entropy shows in the Fig. 3 and Table 3 shows that the homogeneity of the proposed clustering is better than existing methods. The entropy of the proposed work is improved as a minimum of 4% than existing system and it leads to maximum of 20%. The zero homogeneity is also possible in the proposed methods, if the proposed method is trained with more number of documents. The F-measure of the proposed work is shown in the Table 2 and Fig. 2 are shown that the performance is improved as a minimum of 5% than existing system and it leads to maximum of 14%. From these results it is concluded that the proposed MCMM will effective than existing methods.

Therefore the precision and recall are optimal than existing system in the proposed MCMM. From the result, the proposed Meta data conceptual mining model (MCMM) proves that it is an effective process for text clustering. And the proposed MCMM leads to more number of classifications per unit time than existing methods.

REFERENCES

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. DOI: 10.1109/ICTAI.2009.41

Chenghua, L., H. Yulan, E. Richard and R. Stefan, 2012. Weakly supervised joint sentiment-topic detection from text. IEEE Trans. Knowl. Data Eng., 24: 1134-1145. DOI: 10.1109/TKDE.2011.48

Chen, C.C. and M.C. Chen, 2012. TSCAN: A content anatomy approachto temporal topic summarization. IEEE Trans. Knowl. Data Eng., 24: 170-181.

Cai, D., X. He and J. Han, 2011. Locally consistent concept factorization for document clustering. IEEE Trans. Knowl. Data Eng., 23: 902-913. DOI: 10.1.1.188.4136

El-far, M., L. Moumoun, M. Chahhou, T. Gadi and R. Benslimane, 2011. Comparing between data mining algorithms: "Close+, Apriori and CHARM" and "Kmeans classification algorithm" and applying them on 3D object indexing. Proceedings of the International Conference on Multimedia Computing and Systems, Apr. 7-9, IEEE Xplore Press, Ouarzazate, pp: 1-6. DOI: 10.1109/ICMCS.2011.5945722

Dhaliwal, J., S.J. Puglisi and A. Turpin, 2012. Practical Efficient String Mining. IEEE Trans. Knowl. Data Eng., 24: 735-744. DOI: 10.1109/TKDE.2010.242

Koteeswaran, S., P. Visu and J. Janet, 2012. A review on clustering and outlier analysis techniques in datamining. Am. J. Applied Sci., 9: 254-258. DOI:10.3844/ajassp.2012.254.258

Kundu, G., M.M. Islam, S. Munir and M.F. Bari, 2008. ACN: An associative classifier with negative rules. Proceedings of the 11th IEEE International Conference on Computational Science and Engineering, Jul. 16-18, Sao IEEE Xplore Press, Paulo, pp: 369-375. DOI: 10.1109/CSE.2008.48

Mazid, M.M., A.B.M.S. Ali and K.S. Tickle, 2009. A comparison between rule based and association rule mining algorithms. Proceedings of the 3rd International Conference on Network and System Security, Oct. 19-21, IEEE Xplore Press, Gold Coast, QLD, pp: 452-455. DOI: 10.1109/NSS.2009.81

Kantardzic, M., 2003. Data Mining: Concepts, Models, Methods and Algorithms. 1st Edn., Wiley-Interscience, Hoboken, NJ., ISBN-10: 0471228524, pp: 345.

Zhong, N., Y. Li and S.T. Wu, 2012a. Effective pattern discovery for text mining. IEEE Trans. Knowl. Data Eng., 24: 30-44. DOI: 10.1109/TKDE.2010.211

Zhong, N., Y. Li and S.T. Wu, 2012b. Effective pattern discovery for text mining. IEEE Trans. Knowl. Data Eng., 24: 30-44. DOI: 10.1109/TKDE.2010.211

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. DOI: 10.1109/ICCITECHN.2010.5723889

Shehata, S. and M. Kamel, 2010. An efficient concept based mining model for enhancing text clustering. IEEE Trans. Knowl. Data Eng., 22: 1360-1371. DOI: 10.1109/TKDE.2009.174

Li, Y., C. Zhang and J.R. Swan, 2000. An information filtering model on the web and its application in jobagent. Knowl. Based Syst., 13: 285-296.

Barros, R., D. Isidoro and R. Aragues, 2012. Three study decades on irrigation performance and salt concentrations and loads in the irrigation return flows of La Violada irrigation district (Spain). Agric. Ecosyst. Environ., 151: 44-52.