ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 5, Issue 1 (Jan. - Mar. 2017)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2017)

Investigation on Spearman Correlation Based Clustering Technique for Dimensionality Reduction in Web Content Mining

[I]Shanmugapriya M., [II]P Jeyanthirani, [III]Dr R.S.Vetrivel
[I]Research Scholar, Computer Science, Subramanya College of Arts & Science, Palani
[II,III]Assistant Professor, Professor, Computer Science, Subramanya College of Arts & Science, Palani

Abstract
Dimensionality reduction has a great significant and importance in data mining. Spearman clustering is used in exploratory data analysis in data compression, information retrieval and image segmentation. The dimension reduction process is applied in large areas such as data mining, forecasting, image processing and genomic analysis. Spearman correlation coefficient is used to examine integrity of the data structure in original and reduced space. The existing work was introduced a Partially Expected Edit Distance Reduction (PEEDR) and Correlated Probabilistic Graphs Spectral (CPGS) clustering algorithms. Clustering efficiency is increased with the help of designing pruning techniques. In PEEDR method, cluster graph is significantly improved by means of including or extracting vertices from some clusters. However, due to the sparseness in cluster, the correlated probability graph produces the noise and hence difficult to solve. The partially expected edit distance algorithm does not effectively reduce the dimensionality space. In order to overcome the above limitations, Investigation on Spearman Correlation Based Clustering Technique for Dimensionality Reduction in Web Content Mining is designed. This method improves Clustering technique for calculating spearman correlation coefficient. Initially, principal component analysis (PCA) is applied to dimension reduction process for choosing the dimensions with higher variances. Next, spearman correlation is calculated to the two different queries for retrieving more relevant documents. Then, spearman correlation based clustering is applied to clusters the input queries. Finally, correlation filter is employed to separate the highly correlated cluster with dimensionality reduction for retrieving relevant documents in Web mining technique.

Key Terms
Dimensionality Reduction, Spearman Cluster, Correlation Coefficient, Data Mining

I. Introduction

A. Data Mining

Data Mining is a process of extracting useful information in large amount of data using the methods of artificial intelligence, machine learning, statistics and database management systems. Data mining is the study of Knowledge Discovery in Database (KDD). KDD is a process of relating to more than one branch of knowledge in computer science. Data mining software is the analytical tools for analyzing data. It permits users to examine data from various dimensions or angles for classifying and reviewing the recognized connections. Data mining is the method of discovering either connections or patterns with set of fields in huge relational records.

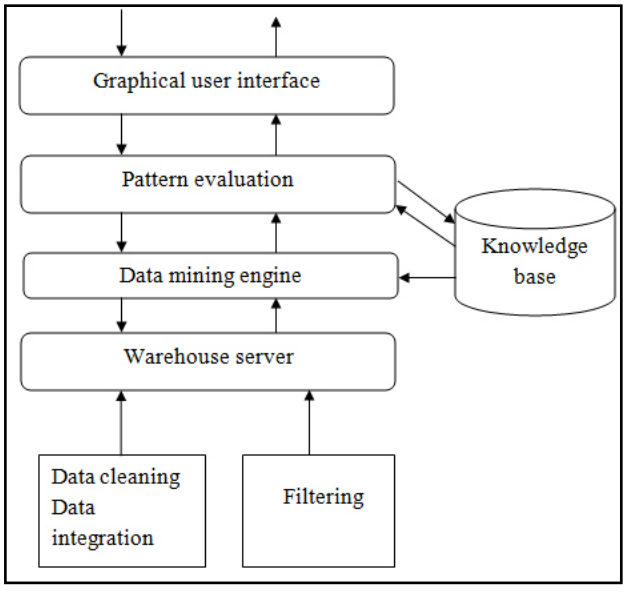

Fig. 1.1 : Typical data mining system

Figure 1.1 shows the architecture of typical data mining system. The main elements of data mining system contains source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base. The data mining system has the sources of database, data warehouse, World Wide Web (WWW), text files and other documents. Data warehouses hold more than one database, document files, spreadsheets or other kinds of information. Sometimes, data may exist even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

The data mining engine is the basic element of any data mining system. It consists of number of components for achieving data mining tasks of association, classification, characterization, clustering, prediction, time-series analysis etc. The pattern evaluation element is mainly responsible for measuring importance of pattern with the help of threshold value. It associates with the data mining engine for concentrating the search towards importance of patterns. The Graphical User Interface (GUI) is used to support the user for using the system simply and effectively without knowing the real complexity behind the process.

When the user identifies a query, GUI interacts with the data mining system to produce the result in a efficient manner. The knowledge base is a significant element for entire data mining process. The data mining engine obtains the inputs from the knowledge base for generating the results in an accurate and reliable manner. The pattern evaluation communicates with the knowledge base to get inputs and also to update it. These different elements of data mining system need to interact properly with each other for executing the complex process of data mining successfully.

B. Web Mining

Data mining is a widely used method for determining significant information in a multipart dataset and for many improvements in

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

the subject of web mining. Web mining is the important operations of data mining method used to identify the patterns from the Web. Web is a selection of inter-related documents on several Web servers. Based on the targets analysis, Web mining is divided into Web content mining, Web structure mining and Web usage mining. Web content mining is the most importance one for extracting relevant information from Web.

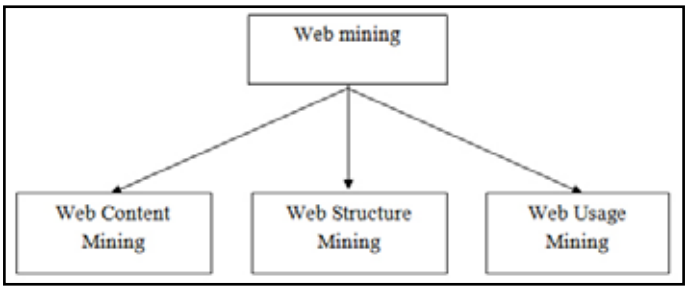

Fig. 1.2 : Types of Web Mining

Figure 1.2 illustrates the classification of web mining. Web mining generally separated into three different kinds depends on the type of data to be mined. Extract "snippets" from a Web document that denotes Web content mining. Web Structure Mining is used to preprocessing the entire web graph or discovers the interesting graph patterns. Web Usage Mining deals with the task such as, user detection, session creation, robot identification and extracting usage path patterns.

1. Web Content Mining

Web Content Mining is the process of retrieve the related information in the contents of Web documents. Web content mining introduces searching effect of relevant information for retrieving exact users quires and terms. Content data includes text, images, audio, video or structured records and so on. Web content mining is an automatic process that exists over keyword extraction. The process of text mining through the Web content was significantly researched. Issues specified in content mining are, topic discovery, retrieving correlation patterns, clustering of web documents and classification of Web Pages.

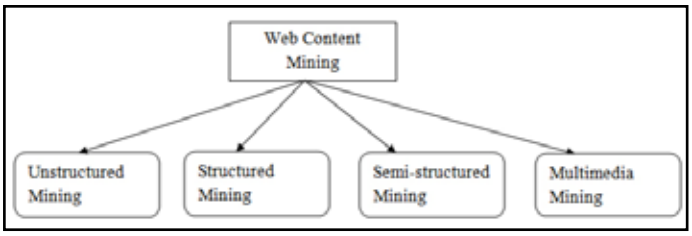

Fig. 1.3 : Techniques of Web Content Mining

Figure 1.3 shows the techniques of Web content mining. Web content mining technique involves huge considerations on data mining and text mining but, all of its techniques are not based on them. Generally, Web content mining is the subfield of data mining but it is does not categories the subfield of text mining. Text mining holds the operations of evaluating unstructured textual data, extract numeric keyword from the text and generate the information accessible to various data mining algorithms. In this case of Web mining, the entire data is does not in the textual form, it also process the non textual data such as Web server logs and transaction based data.

The Web content data in unstructured or structured or semi structured form like free text, data in the tables and HTML documents respectively. Different techniques are required to be applied in all three cases. The concept of web content mining compresses the techniques and methods for summarizing, categorization and clustering of the web page contents. It generates useful and interesting patterns of user requirements and contribution formats. Web content mining concentrates on the knowledge detection from web pages and collecting text and multimedia documents like graphics, audios, videos and animations. It is generally depended upon information extraction and text mining like, text classification, clustering and the information visualization.

C. Clustering Analysis

Cluster analysis or clustering is the process of combining a set of objects that are in same group (cluster) or more identical in other groups (clusters). It is a major task for investigating the data mining and a general method for statistical data analysis that applied in machine learning, pattern recognition, image analysis, information retrieval bio informatics and so on.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It is completed by various algorithms that differ significantly due to the compression of different set of objects. General notions of clusters contains group through small distances between the cluster parts, intense areas of the data space, intervals or specific analytical distributions. Hence, clustering techniques solves multi-objective optimization complexity. The applicable clustering algorithm and metric selections are based on the separate data set and intended use of the results. Some of the cluster analyzes are does not process automatically, but a repeated process of knowledge discovery that holds trial and failure. It frequently requires for adjusting data preprocessing and model metrics until the result attains the preferred properties.

In addition, the clustering process includes number of terms with same meaning, containing automatic classification, numerical taxonomy and topological analysis. This in turns misunderstanding between researchers happen from the fields of data mining and machine learning technique since they utilize the similar terms and same algorithms, but have different goals.

II. Literature Survey

1. Multitask Spectral Clustering by Exploring Inter task Correlation

In this paper [1] a novel clustering model called multitask spectral clustering presented to discover the specific characterizes of cluster label matrix. There are two types of correlations are considered is method.

Initially, inter task clustering correlation is considered to verify the relations between diverse clustering tasks. Next, intra task learning correlation is considered to allow the operation of learning cluster labels and learning mapping function to support each other. Based on the general low-dimensional assumption, a novel l2,p- norm regularize is integrated for organizing the coherence of the entire tasks. In addition, mapping properties of a cluster label matrix is used an explicit mapping function that correspondingly forecast the cluster labels for each individual process.

2. TotalPLS: Local Dimension Reduction for Multi-category Microarray Data

This paper [2] introduced the PLS-based feature selection algorithm (PLSRFE) for multi-category classification. Then, a new local dimension reduction algorithm namely TotalPLS is

executes an information fusion of PLS-based feature selection and feature extraction in an integrated PLS framework.
At first, PLS-based recursive feature elimination (PLSRFE) in multi-category problems are derived. A novel reduction algorithm enhances the recognition precision, interpretability and visualization for extracting the potential structure to high-dimension multi-category data. Furthermore, the algorithm is effectively used to microarray data for analyzing co-expression and co-regulation.

3. Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence

In this paper [3], the integration of cluster data called high-dimensional similarity based PCM (SPCM) is designed with ant colony optimization intelligence algorithm. PCM is an essential for clustering unstructured data without knowing knowledge about cluster number from the user.
The PCM was developed into similarity based by applying mountain method with it. This is sufficient clustering is used to develop the optimization by means of ant colony algorithm with swarm intelligence. Therefore, the synthetic data sets are applied to achieve the scalable clustering technique and check the estimation results.

4. A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique

This paper [4] developed a hybrid clustering algorithm that depended upon similarity in the structure of time series data. Initially, based on the similarity time the time series data are clustered into subfield cluster.
Then, based on similarity in form the subfield clusters are combined with the help of k-Medoids algorithm. This model has two operations. It is higher accurate compared to traditional and hybrid methods and it identifies the similarity in structure between time series data with a low complexity. This method is experienced broadly by using semantic and real-world time series datasets for calculating the precision of the clustering algorithm.

5. Dimensionality Reduction by Weighted Connections between Neighborhoods

In this paper [5], a dimensionality reduction method was developed by weighted associations among neighborhoods.
A dimensionality reduction method is employed to enhance K-Isomap method for tries to maintain the associations between neighborhoods in dimension reduction process. The strength of this reduction technique was examined by three classic examples which are generally used in the algorithms based on manifold. In addition, a development of K-Isomap algorithm was applied to reduce the dataset into high-dimensional space to low-dimensional space.

6. A Kernel Based Neighborhood Discriminant Submanifold Learning for Pattern Classification

This paper [6] presents a kernel neighborhood specific analysis based on the supervised kernel improvement of locality preserving projection.
It is nonlinearly maps the original data into a kernel space in which two graphs are generated to place a within submanifold and between-class submanifold. Then it reduces the computation between the within-class representation and the between-class representation of the submanifolds is considered to divide each submanifold produced by each class. The main objective of kernel neighborhood discriminant analysis is used to improve the submanifold based algorithm to a general model that organizes a given object to a predefined class efficiently.

7. Fuzzy C-Means and Cluster Ensemble with Random Projection for Big Data Clustering

In this paper [7], a new fuzzy c-means clustering algorithm with random projection is integrated for theoretical analysis.
Empirical solutions are proves that the new algorithm efficiently maintains the accuracy of original FCM clustering. It also provides more efficient than original clustering and clustering among singular value decomposition. As a result, a new cluster collection approach based on FCM clustering with random projection is also introduced. The new aggregation method successfully compute the spectral embedding of data with cluster centers based representation which scales linearly with data size.

8. Regularized Embedded Multiple Kernel Dimensionality Reduction for Mine Signal Processing

This paper [8], describes a new multiple kernel dimensionality reduction method for mine signal processing. It avoids SDP relaxation and also increases performance of multiple kernel dimensionality reduction by considering EGE and regularized trace quotient maximization.
In addition, this method produces valuable use of the binary search and different optimization design for successfully deriving optimal solutions. Multiple kernel learning into novel graph embedding is applied to improve the of single kernel dimensionality reduction operation.

9. Spectral Nonlinearly Embedded Clustering Algorithm

In this paper [9], common spectral embedded method was designed to combine the true cluster transfer matrix for high-dimensional data within a nonlinear extension by a predefined embedding function.
Based on this framework, different algorithms are provided by several embedding functions. This function aims at learning the final cluster assignment matrix and transferring the low dimensionality space concurrently. Moreover, the spectral embedded framework logically solves the out-of-sample development issues.

10. Parallel rare term vector replacement: Fast and effective dimensionality reduction for text

This paper [10], presents a dimensionality reduction approach for text, along with a parallel algorithm appropriate to confidential memory similar computer systems.
Based on the Zipf's law, the common of indexing terms happens in a minimum number of documents. Dimensionality reduction algorithm returns infrequent terms by calculating a vector which states their semantics in terms of frequent terms. This process generates a projection matrix used to a amount matrix and individual document and query vectors.

11. A three-stage unsupervised dimension reduction method for text clustering

In this paper [11], presents a three-stage dimension reduction models for reducing unrelated, redundant and noisy features in the original search space without loss of useful information. These

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

models contain the benefits of the FS and the FE methods for generating low dimension feature subspace.
Moreover, the collection of feature subsets and operation of the clustering algorithm highly based on the parameters values. The primary cluster centroid selection concerns the efficiency of the clustering algorithm extensively.

12. Enhancing density-based clustering: Parameter reduction and outlier detection

This paper [12], presented a simple and efficient algorithm that able to enhance the execution of density-based clustering maintaining the asymptotic running time of DBSCAN.
It is based on the theory of space stratification and it efficiently identifies the diverse densities in the dataset. This in turns, the objects of the original space is ranked. It develops such knowledge by designing the original data into a space with one more dimension. It operates a density based clustering that considered the account of reverse-nearest-neighbor of the objects.

13. A Novel Ranking-Based Clustering Approach for Hyper spectral Band Selection

In this paper [13], a fast density-peak-based clustering algorithm was developed to for hyper spectral band range. The ranking score of each band is calculated by weighting the normalized local density and the intra-cluster distance quite than similarly considered them into account.
An exponential-based learning rule is applied to modify the cutoff threshold for a several number of preferred bands. This clustering approach is also improves the performance of the system. Furthermore, an effective plan called, isolated-point-stopping criterion is enhanced to robotically calculate the appropriate number of bands to be selected. The effective clustering process is applied to prevent the appearance of an isolated point (the only point in one cluster).

14. Document clustering method using dimension reduction and support vector clustering to overcome sparseness

This paper [14], presents a clustering method for answering three problems in document clustering.
A combined clustering method using dimension reduction and K-means clustering based on support vector clustering and outline measures are constructed in this method. Moreover, the sparseness in patent document clustering problem is tries to solve in this process. Here, the documents are exchanged to structured data for document clustering.

15. Geodesic distance based fuzzy c-medoid clustering – searching for central points in graphs and high dimensional data

In this paper [15], an algorithm for mining central objects in graphs and high dimensional data were designed to minimize the computation cost.
The modified fuzzy c-medoid clustering algorithm is applied to calculate the shortest path distance and chooses potential cluster centers between the set of most essential objects. The fuzzy c-medoid clustering algorithm also holds the data that lies on a low dimensional extension of a high dimensional feature space. The convergence of the algorithm is enhanced with the help of selecting the centrality measure properly the substantially.

Investigation on Spearman Correlation Based Clustering Technique for Dimensionality Reduction in Web Content Mining

Spearman clustering is used to investigative the data analysis process of data compression, information retrieval and extraction in data mining technique. Dimensionality reduction is the most important process in Web content mining for reducing the dimensionality of irrelevant data. Spearman correlation is employed to examine integrity of the data structure in original and reduced space. But, most of the techniques are does not reduce the dimensionality effectively.
Spearman Correlation Based Clustering (SCBC) technique is introduced for dimensionality reduction in web content mining. Initially, Principal Component Analysis (PCA) is applied to extracting uncorrelated data. PCA is involved in dimension reduction process for selecting the dimensions with higher variances. Then, spearman correlation is measured to the two different input queries (data's) for retrieving most relevant documents. Here, the spearman correlation between two queries (data) will be high then the spearman correlation co-efficient of high correlation is denoted as positive value.
On the other hand, low correlation between the data represents the negative value. Next, spearman correlation based clustering is applied to clusters the correlated and uncorrelated data. Finally, correlation filter is employed to filter the high and low correlated clusters for reducing the dimensionality of low correlation (irrelevant data) clusters in Web content mining. As a result, more relevant document (data) is retrieved with dimensionality reduction (irrelevant data space) in Web content mining technique.

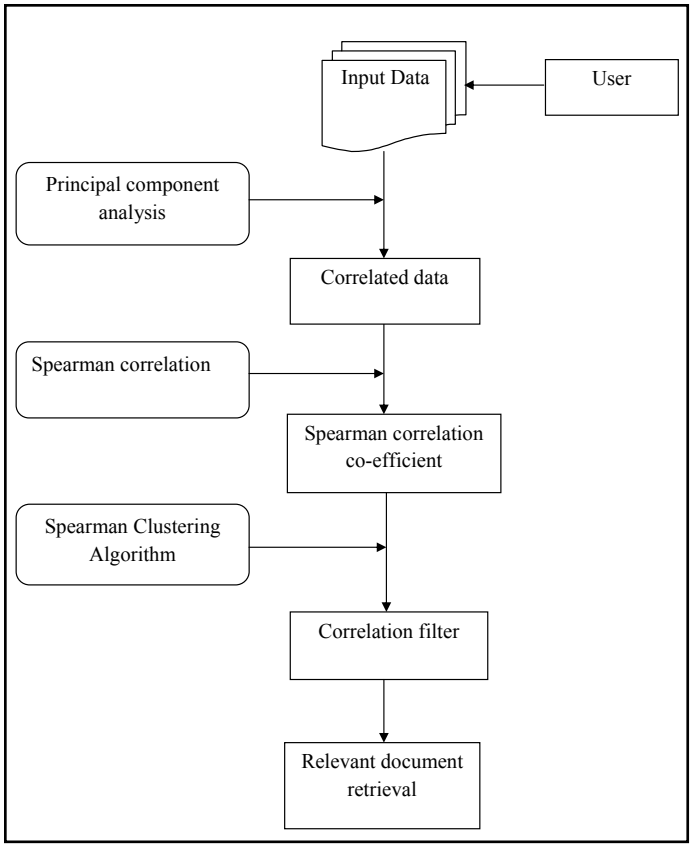

Fig. 3.1 : Architecture Diagram of Spearman Correlation Based Clustering Technique

Figure 3.1 shows the architecture of spearman correlation based clustering technique in Web content mining. At first, uncorrelated

data is extracted with the help of Principal component analysis. Then, correlation co-efficient is calculated by using spearman correlation co-efficient. Next, Spearman based correlation co-efficient clustering algorithm is applied to cluster the correlated and uncorrelated data. Finally, relevant document or data is retrieved by filtering high correlation data.

The spearman correlation based cluster scheme process is divided into:

a) principal component analysis
b) Spearman correlation
c) Correlation of clustering
d) Correlation filter

a) Principal component analysis

The querying and mining of correlations become highly demanding to redefine statistical data mining problems. In data mining technique, dimensionality reduction is the most important process for retrieving the most relevant information from the Web. Principal component analysis (PCA) based dimension reduction is employed to chooses the dimensions with higher variances. In general, high dimensional data are transferred into low dimensional data with the help of principal component analysis (PCA) where coherent patterns are identified efficiently. This kind of dimension reduction process is applied in large areas such as data mining, forecasting, image processing and genomic analysis. Mathematically, PCA is applied to detect the capital low rank estimation (in L2 norm) of the data through the singular value decomposition. Based on the two characteristics, the input data (documents) is extracted with the help of PCAs. Initially, based on the maximum amount of total variance in the observed data (query) is extracted by using PCA.

Next, based on the maximum amount of total variance in the data set is extracted, since it does not considered by the initial one. The remaining data is extracted by using above two characteristics. Each data considered to the maximal amount of variance in the observed data's that does not assumed by preceding components. When the principal component analysis is completed, the result shows varying degrees of correlation with the observed data. Principal components are calculated using the Eigen value decomposition of a data covariance matrix. Covariance matrix is chosen when the variances of queries are very high compared to correlation. Therefore, uncorrelated data is effectively extracted for reducing the search space or dimensionality in data mining process. The PCA is employed on the original data set for attaining reduced dataset that including possibly uncorrelated variables for improving the efficiency.

b) Spearman correlation

Spearman correlation is used to examine goodness of the data structure in original and reduced space. Spearman correlation coefficient is applicable for continuous and discrete variables, comprising ordinal variables. There are no frequent query values are occurred in the spearman correlation therefore, accurate spearman correlation coefficient of +1 or −1 is appeared when each of the data are perfect monotone function of the other. The relevant query is extracted to calculate the correlation among the queries as follows. Intuitively, the spearman correlation among two queries are high when observations include a related (or identical for a correlation of 1) rank between the two queries. Otherwise, low correlation include a dissimilar (or fully opposed for a correlation of -1) rank between the two queries. This in turns, high correlation queries are extracted and grouped to reduce the dimensionality with the help of spearman correlation coefficient.

Moreover, a value of 1 represents a linear equation that defines the relationship between X and Y more accurately, with all data points lying on a line for which Y increases as X increases. A value of −1 represents the all data point's lies on a line for which Y decreases as X increases. A value of 0 represents there is non-linear correlation among the variables. Additionally, note that (Xi − X) (Yi − Y) is positive if and only if Xi and Yi lie on the equal side of their individual means. As a result, the correlation coefficient is positive if Xi and Yi leads to concurrently larger than or concurrently lesser than their corresponding terms. The correlation coefficient is negative if Xi and Yi leads to lie on opposite sides of their corresponding terms.

c) Correlation of clustering

Here, high correlated queries are collected for grouping process by suing clustering technique. Clustering is the process of identifying group of objects that are identical to one another and diverse from the objects in other groups. Dimensionality reduction is used to transforms the high-dimensional data for retrieving most relevant data in terms of reducing search space. A dimension is described as the measurement of a certain aspect of an object. Dimensionality reduction is the process of reducing the dimensionality of an object for performing pattern recognition, machine learning, document extraction and data mining. The main aim of dimensionality reduction is used to avoid irrelevant and unnecessary data for increasing the quality of data in data mining technique.

d) Correlation filter

The correlation filter is applied to calculate the high and low correlation among the clustering. It is used to easily identify the high correlated queries for retrieving most related documents. Due to the word of caution and sensitive scale correlation, correlation filter is necessary for a significant correlation comparison. Data columns with more parallel trends are also possible to execute more related information. In addition, the convex relation on spearman clustering obtains an optimal correlated clustering. Convex optimization formulation is the weighted form of low-rank matrix decomposition and address correlation among the queries. The optimal correlated cluster improves the accurate clustering and achieves under less restrictive condition. The optimization of cluster includes a small number of errors. The optimization reduces the weighted grouping of the two types of errors.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Input: Data set that containing text documents, Input query with number of terms
Output: Relevant document retrieval
Step 1: Begin Step 2: Let each document in data set is separated into correlated and uncorrelated data Step 3: Measure of correlation co-efficient based on the spearman correlation Step 4: Find the pair of clusters based on the spearman correlation clustering Step 5: Filter the high correlation cluster Step 6: Repeat merging process until obtain a most relevant data Step 7: End

Fig. 3.2 : Algorithm of Spearman Correlation Based Clustering

IV. Results and Discussion

A. Performance Metrics

This work quantifies the performance of spearman correlation based clustering technique by comparing it to partially expected edit distance reduction (PEEDR) in vertices adding or removing from some clusters. It evaluates the performance of the methods by evaluating the clustering efficiency, error rate and cluster generation time, when compared to the existing system. This system performs and analyzed the metrics in java platform by:

- Clustering efficiency
- Error rate
- Cluster Generation Time

1. Clustering efficiency

Clustering efficiency is measured for spearman correlation clustering in proposed SCBC method where clusters are formed based on the semantic similarity of given words from different text documents. Clustering efficiency is measured in terms of percentage (%). If clustering efficiency is high, then the method is said to be more efficient for text classification.

Table: 4.1. Tabulation for clustering efficiency

Number of input queries	Clustering efficiency (%)	
	EECM (Existing)	SCBC (Proposed)
10	60.9	74.6
20	63.2	76.9
30	66.1	79.1
40	67.3	83.2
50	70.8	84.7

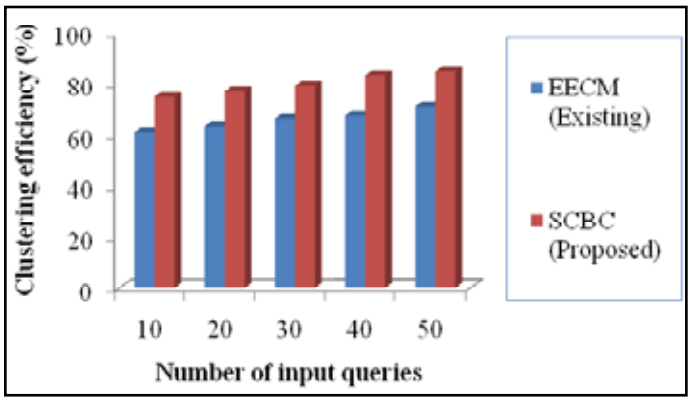

Fig. 4.1 : Measure of clustering efficiency

Figure 4.1 demonstrates clustering efficiency (%). X axis represents the number of input queries whereas Y axis denotes clustering efficiency using both the existing Effective and efficient clustering methods (EECM) and Spearman Correlation Based Clustering (SCBC). When the number of input queries increased, clustering efficiency also increased accordingly.

Figure shows better performance of spearman correlation clustering in terms of than existing system. Spearman correlation based clustering technique for dimensionality reduction in web content mining improves the clustering efficiency by 21% when compared to existing system.

2. Error rate

Error rate is defined as the measure of error observed during the clustering of unstructured text documents. Error rate is measured in terms of percentage (%). If error rate is low, then the method is said to be more effective.

Table: 4.1 : Tabulation for error rate

Number of input queries	Error rate (%)	
	EECM (Existing)	SCBC (Proposed)
10	26	18
20	28	20
30	31	24
40	35	25
50	37	29

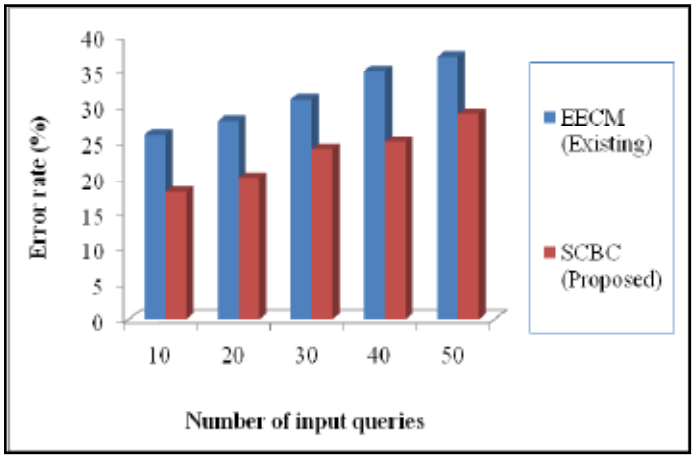

Fig. 4.2 : Measure of error rate

Figure 4.2 demonstrates Error rate (%). X axis represents the

number of input queries whereas Y axis denotes error rate using both the existing Effective and efficient clustering methods (EECM) and Spearman Correlation Based Clustering (SCBC). When the number of input queries increased, error rate gets decreased accordingly.
Figure shows better performance of spearman correlation clustering technique in terms of than existing system. Spearman correlation based clustering technique for dimensionality reduction in web content mining reduces the error rate by 26% when compared to existing system.

3. Cluster Generation Time

Cluster generation time is measured by time taken to cluster the text documents from large unstructured dataset with respect to total number of input queries. Cluster generation time is measured in terms of seconds (Sec). When generation time is lower, the method is said to be more efficient.

Table: 4.3 : Cluster Generation Time

Number of input queries	Cluster Generation Time (Sec)	
	EECM (Existing)	SCBC (Proposed)
10	45	43
20	52	49
30	57	54
40	62	59
50	67	64

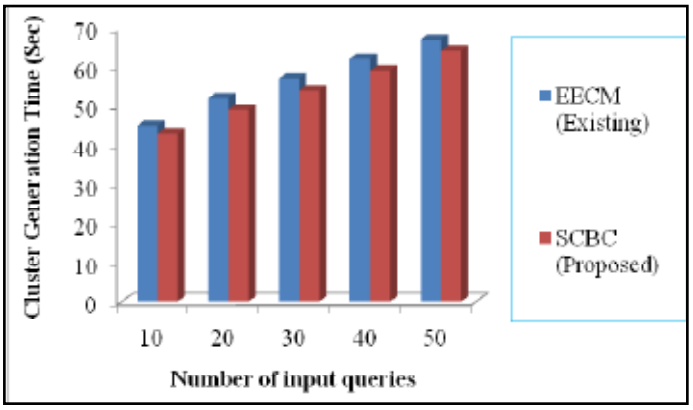

Fig. 4.3 : Cluster Generation Time

Figure 4.3 demonstrates cluster generation time (sec). X axis represents the number of input queries whereas Y axis denotes cluster generation time using both the existing Effective and efficient clustering methods (EECM) and Spearman Correlation Based Clustering (SCBC). When the input queries increased, cluster generation time gets decreases accordingly.
Figure 4.3 shows better performance of spearman correlation clustering technique in terms of time than existing system. Spearman correlation based clustering technique for dimensionality reduction in web content mining reduces the cluster generation time by 5% when compared to existing system.

V. Conclusion

Conclusion and Future work

Finally conclude a spearman correlation clustering, effectively reduces the dimensionality in web content mining. Based on the correlation clustering technique, variance among the queries is measured. Principal component analysis (PCA) is applied to dimension reduction process for selecting the dimensions with higher variances. In addition, spearman correlation based clustering is used to clusters the input queries. Moreover, correlation filter is applied to separate the highly correlated cluster with dimensionality reduction for extracting relevant documents in Web mining technique.

References

[1] *Yang Yang, Zhigang Ma, Yi Yang, Feiping Nie, and Heng Tao Shen, "Multitask Spectral Clustering by Exploring Intertask Correlation", IEEE Transactions on Cybernetics , Volume 45, Issue 5, May 2015 , Pages 1083 – 1094.*

[2] *Wenjie You, Zijiang Yang, Mingshun Yuan, and Guoli Ji, "TotalPLS: Local Dimension Reduction for Multicategory Microarray Data", IEEE Transactions on Human-Machine Systems, Volume 44, Issue 1, February 2014, Pages 125-138.*

[3] *Thenmozhi Srinivasan and Balasubramanie Palanisamy, "Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence", Hindawi Publishing Corporation, The Scientific World Journal, Volume 2015, April 2015, Pages 1-5.*

[4] *Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, Hamid A. Jalab, Mohammad Amin Shaygan, and Alireza Jalali, "A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique", Hindawi Publishing Corporation, Te Scientific World Journal, Volume 2014, March 2014, Pages 1-13.*

[5] *Fuding Xie, Yutao Fan, andMing Zhou, "Dimensionality Reduction by Weighted Connections between Neighborhoods", Hindawi Publishing Corporation, Abstract and Applied Analysis, Volume 2014, April 2014, Pages 1-6.*

[6] *Xu Zhao, "A Kernel Based Neighborhood Discriminant Submanifold Learning for Pattern Classification", Hindawi Publishing Corporation, Journal of Applied Mathematics, Volume 2014, February 2014, Pages 1-11.*

[7] *Mao Ye, Wenfen Liu, JianghongWei, and Xuexian Hu, "Fuzzy □-Means and Cluster Ensemble with Random Projection for Big Data Clustering", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2016, June 2016, Pages 1-14.*

[8] *Shuang Li, Bing Liu, and Chen Zhang, "Regularized Embedded Multiple Kernel Dimensionality Reduction for Mine Signal Processing", Hindawi Publishing Corporation, Computational Intelligence and Neuroscience, Volume 2016, April 2016, Pages 1-13.*

[9] *Mingming Liu, Bing Liu, Chen Zhang, andWei Sun, "Spectral Nonlinearly Embedded Clustering Algorithm", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2016, June 2016, Pages 1-10.*

[10] *T. Berka , M. Vajteršic, "Parallel rare term vector replacement: Fast and effective dimensionality reduction for text", Elsevier, Journal of Parallel and Distributed Computing, Volume 73, Issue 3, March 2013, Pages 341–351.*

[11] *Kusum Kumari Bharti, P.K. Singh, "A three-stage unsupervised dimension reduction method for text clustering", Elsevier, Journal of Computational Science, Volume 5, Issue 2, March*

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

2014, Pages 156–169.

[12] Carmelo Cassisi,AlfredoFerro n, Rosalba Giugno,Giuseppe Pigola, Alfredo Pulvirenti, "Enhancing density-based clustering: Parameter reduction and outlier detection", Elsevier, Information Systems, Volume 38, Issue 3, May 2013, Pages 317–330.

[13] Sen Jia, Guihua Tang, Jiasong Zhu, and Qingquan Li, "A Novel Ranking-Based Clustering Approach for Hyperspectral Band Selection", IEEE Transactions on Geoscience And Remote Sensing, Volume 54, Issue 1, January 2016, Pages 88-102.

[14] Sunghae Jun , Sang-Sung Park , Dong-Sik Jang, "Document clustering method using dimension reduction and support vector clustering to overcome sparseness", Elsevier, Expert Systems with Applications, Volume 41, Issue 7, 1 June 2014, Pages 3204–3212.

[15] AndrásKirálya, ÁgnesVathy-Fogarassyb, JánosAbonyia, "Geodesic distance based fuzzy c-medoid clustering – searching forcentral points in graphs and high dimensional data", Elsevier, Fuzzy Sets and Systems, Volume 286, 1 March 2016, Pages 157–172.