Application of K-means Algorithm to Web Text Mining Based on Average Density Optimization

FAN Guang-Ling[1], LIU Yu-Wei[2], TONG Jan-Qiang[3], ZHAO Sheng-Hai[3], NIE Zhi-Quan[4]
[1] School of Mathematics and Statistics, Northeast Petroleum University
Daqing, 163318, China
[2] School of Petroleum Engineering, Northeast Petroleum University
Daqing, 163318, China
[3] Daqing Oilfield Powerlift Pump Industry Co., Ltd
Daqing, 163311, China
[4]Dar Petroleum Operating Company
Juba, 81111, South Sudan
fanguangling@126.com

Journal of Digital Information Management

ABSTRACT*: Text information is increasing at an explosive speed with the advent of the Internet. However, this situation has given rise to the problem of abundant information with relative deficiency of knowledge. Therefore, finding a way to seek target information rapidly and accurately has become a research hotspot. This study presented a method to improve web text clustering accuracy and integrity. First, the dk-means algorithm was modified, and the k-means algorithm based on average density optimization was proposed. Second, a web text clustering model was designed, and in-depth research on the key technology of web text clustering was conducted. Finally, the k-means algorithm based on average density optimization (Adk-means algorithm) was applied to the web text clustering model, and clustering and classification of web text were completed. Experiment showed that the purity and mutual trust values of the Adk-means algorithm are higher than those of the dk-means algorithm, and the modified algorithm is greatly improved in terms of accuracy, integrity, and performance of partitioned clusters. When clustering text, the Adk-means algorithm has high polymerization and similarity within classification. Research results were applicable to text clustering. When used in Internet text searching, the Adk-means algorithm is a highly efficient information retrieval technology that can improve searching speed and accuracy.*

Subject Categories and Descriptors
K.2.8 [Database Applications]: Data mining; **B.2.4 [High-Speed Arithmetic]:** Algorithms

General Terms: Computational Swarm Intelligence, Clustering, Data mining, Classfication

Keywords: Web Text Clustering, Word Segmentation, Eigenvector, dk-means Algorithm, Adk-means Algorithm

Received: 5 September 2015, Revised 1 October 2015, Accepted 6 October 2015

1. Introduction

The Internet has been rapidly developing in recent years, and increasing numbers of organizations, groups, and individuals are delivering and searching information. Finding accurate information that people want is time and energy consuming, and difficult. One approach to this problem is web mining. K-means algorithm based on partition is a research hotspot in web mining. Local and foreign researchers have conducted extensive research to develop schemes to improve this algorithm. However, these improved methods still retain some inherent defects of the original algorithm. This study mainly improved the dk-means algorithm and proposed the concept of average density. This study partitioned the isolated points of data concentration when the algorithm began, searched among sets that were greater than the average density for selecting clustering centers, and finally distributed the isolated points to the clustering center closest to itself until complete classification of datasets was achieved. The results are helpful for utilizing World Wide Web resources and assisting users to accurately find the data

they need. This study also helps users to save retrieval time and improve the utility value of web documents.

2. Literature Review

In terms of clustering analysis algorithms, local and foreign experts and scholars have conducted numerous studies, most of which focus on improving the traditional k-means algorithm. For example, for displacement between two iterations in the application center, a comprehensive method that is faster than the traditional k-means algorithm is proposed, and it can be applied to medical diagnosis and customer analysis [1–2]. The simple k-means clustering and information classification algorithms are applied to a cloud system [3]. The k-means clustering algorithm itself is a global optimization problem, and its target function has multiple local minimum values with only one global minimum value; this algorithm easily gets stuck in the local minimum iteration process. A new l-means clustering algorithm based on dynamic mining system is proposed, and this algorithm can skip over local minimum value points and obtain a good initial point to enter the dynamic optimization process [4–5]. The algorithm accuracy of the program executed in different input data points is studied, and an algorithm is proposed, which can make the k-means algorithm highly effective and efficient to obtain considerable clustering and reduce complexity [6-7]. An effective method is suggested to address the difficulty in selecting the k value and the initial clustering center. According to the important features of a complicated network, a network model of failure data is established, and the maximum value of different partitioning results is the same as the k value; important nodes are selected as the initial clustering center by calculating the correlation degree of complicated network nodes [8–9]. Given that document data are unsuitable for statistic analysis and machine-learning method, directly analyzing the document is difficult, and document data need to be transformed into structural data. However, structural data are sparse, thereby making them difficult to analyze. A joint clustering method is established by using dimension reduction k clustering based on support vector clustering and profile measurement, and this method solves the problem of sparse document clustering. This method is applied to the news data of a machine learning library in the Irvine Branch of the University of California [10–11]. A new fingerprint image segmentation algorithm called k-means clustering algorithm of dynamic particle swarm optimization (DPSOK) is proposed.

Experimental results show that the DPSOK algorithm can effectively improve the global searching function of k-means clustering [12]. The relevant concept of transactional database clustering is defined, and a clustering strategic transaction database based on k-means is proposed [13]. An updating method of obtaining a dataset with suitable association rules is proposed, and this method can synthesize data evaluation and explanation [14]. An effective in-cluster algorithm is proposed, which can update the random step length of apothem increment and save operation time [15].

Most research from both home and abroad emphasizes theoretical research on algorithms, while a few studies focus on the application of web text mining and information retrieval. The existing k-means algorithm based on the initial clustering center of density optimization has problems, such as a large searching scope, long consumption time, and sensitivity of clustering results to isolated points. A k-means algorithm (Adk-means) based on average density optimization and initial clustering center is proposed to address these problems. This algorithm has a fast convergence rate, strong stability, and high clustering precision, and it eliminates the sensitivity of clustering results to isolated points.

The remaining sections of this paper are organized as follows: Section Two presents the literature review, concludes the research status, analyzes the deficiencies in research, and points out the specific key problems solved by this study. Section Three describes the preprocessing and clustering modules and the process of the Adk-means algorithm. Section Four presents the experimental measurement and evaluates the performance of the Adk-means algorithm. Section Five concludes this study.

3. Methodology

The system model is divided into three parts. The first part is the text preprocessing module that is used to complete data preprocessing, word segmentation, selection of eigenvalues, and construction of eigenvectors. The second part is the text clustering module that acquires the clustering center through the text eigenvectors obtained in the first part. The third part uses the clustering results in the second part to classify the given corpus.

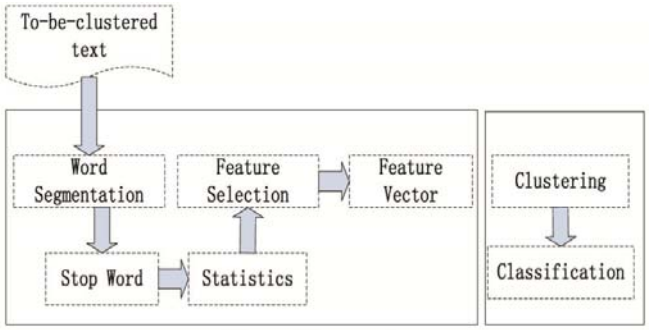

Figure 1. System flow

3.1 Preprocessing Model

(1) Constructing an appropriate stop word list is a key step in the text processing process. The stop word list used in this thesis is the commonly used stop word list downloaded from the Internet, and the text after segmentation uses this list to filter stop words.

(2) Chinese Word Segmentation: This study uses the

ICTCLAS word segmentation system to conduct word segmentation on the obtained corpus. The main function of this system is completing the recognition of Chinese word segmentation and new words, the annotation of parts of speech, and the recognition of named entities. This system was researched and developed by the Chinese Academy of Sciences. The ICTCLAS word segmentation system also supports user dictionary, and the separator is a blank space by default.

(3) Feature Selection: This study adopts document frequency (DF), which is a commonly used dimension reduction method. DF refers to the specific frequency that one feature item presents in multiple documents. The DF algorithm principle is that the DF of one feature item in the given training set is calculated, and some feature items with a low DF are eliminated according to the predefined threshold value. One feature item t appears in documents with a low frequency, and it has no effect on classification. This feature is deleted, and the dimension is reduced. This method can be used in any corpus with a rapid calculation velocity. The time complexity of DF has a linear correlation with the number of documents; hence, this method is applicable to feature selection in large-scale text datasets.

(4) Weight Calculation: The importance degree of each feature item in documents is calculated. This study adopts the term frequency–inverse document frequency (TF-IDF) calculation formula, which is expressed as follows:

$$W(t,\vec{d}) = \frac{tf(t,\vec{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in d} [tf(t,\vec{d}) \times \log(N/n_t + 0.01)]^2}}$$

whereas, $W(t,\vec{d})$: weight of word t in $\vec{d}$; $tf(t,\vec{d})$: frequency of word t in $\vec{d}$; N: total number of samples; n_t: number of documents where word t appears; the denominator in the calculation formula is the normalization factor.

(5) Establishment of Eigenvectors: Vector property is a feature word selected by the DF method. A vector space model (VSM) is saved in the form of a matrix, and the rows represent the document vectors, that is, each row represents one document.

3.2 Clustering Model

The clustering model is as follows:

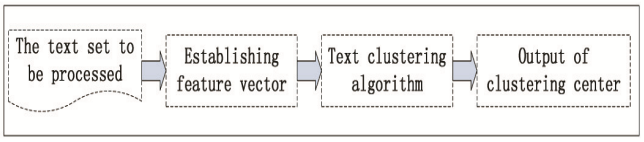

Figure 2. Clustering model

Step 1: The text that needs to be clustered is read into the internal storage.

Step 2: The text is preprocessed, word segmentation is conducted, and the TF-IDF formula is used to calculate the weight of each feature word in each article according to the word frequency of each document.

Step 3: Eigenvectors are established. Then, the VSM of document sets is saved in the form of a matrix. Matrix row refers to one document vector and represents one article.

Step 4: The Adk-means algorithm is used to output the clustering center.

3.3 Text Clustering Model

A document vector model is established. This model is used to calculate the cosine similarity between a document vector and a clustering center. The document is then partitioned to the clustering center that is most similar to it according to the size of cosine similarity, and the article classification is finally completed.

3.4 Motion Constrained Importance Resampling

Basic definition

We set $D=\{ x_i \mid x_i \in R^p, i=1,2\ldots,n\}$, and p is the number of dimensions.

Definition 1: Distance between two sample objects

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T (x_i - x_j)} \quad (1)$$

Definition 2: Average distance between sample points in the dataset

$$MeanDist = \frac{1}{C_n^2} \times \sum d(x_i - x_j) \quad (2)$$

N: total number of sample points, C_n^2: number of sets of the two samples randomly extracted from n sample sets.

Definition 3: Density parameter of samples

$$Dens(x_i) = \sum_{j=1}^{n} u(MeanDist - d(x_i - x_j)) \quad (3)$$

represents the density parameter of sample x_i, where $u(z) = \begin{cases} 1, x \geq 0 \\ 0, \text{其他} \end{cases}$ calculates the density parameters of all samples in D.

Definition 4: Average sample density

$$MDens(x_i) = \frac{1}{n} \sum_{i=1}^{n} Dens(x_i) \quad (4)$$

Definition 5: Isolated points: samples in set D; if $Dens(x_i) < \alpha * MDens(x_i)$, then this point is called an isolated point, where $0<\alpha<1$.

Definition 6: Square error criterion function

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2 \quad (5)$$

E: sum of the square errors of samples; x: data samples that belong to class C_i; m_i: average value of samples in class C_i.

In the Adk-means algorithm, the set is divided into two parts according to the conditions of isolated points. The points that satisfy the condition of isolated points are included in set Noise, other points are placed in another set U, and the density parameters that correspond to U are placed in set T. To find the maximum density parameter in set T, as described in Step 4 of the dk-means algorithm, $T = S \cup S_1$, where S is the set with a density parameter greater than the average density, and S_1 is the density parameter set between the average density and the density of isolated points. In the dk-means algorithm, whenever a clustering center is selected, a global search is conducted in set T. The searching scope greatly increases the searching time. To address the problem, in the dk-means algorithm, we propose a solution, that is, the initial clustering center is selected on the basis of average density. The average density of the sample sets is calculated first. In accordance with the definition of isolated points, isolated points are partitioned to set Noise. In accordance with the solved average density, the parameters that are greater than the average density are placed in set S, and the initial clustering center is selected in set S. The situation in which the maximum value of density parameters is non-unique is processed. This method reduces the searching scope and time while preventing an isolated point from being selected as clustering center.

The algorithm idea is as follows:

Input: sample set $D=\{x_1, x_2, \ldots, x_n\}$, where $x_i=\{x_{i1}, x_{i2}, \ldots, x_{ip}\}$, i=1,2…n; k: number of cluster;

Output: k clustering centers

Steps:

(1) The distance between two samples and the average distance of the entire set are calculated according to Formulas (1) and (2).

(2) The sample density and the average density of the entire set are calculated according to Formulas (3) and (4).

(3) The samples that satisfy the conditions of isolated points are placed in set Noise according to Formula (5), whereas other samples are placed in set U.

(4) Step 2 is executed for set U, and the density parameters that are greater than the average density are placed in set S.

(5) The maximum value S_i of density parameters is found in set S, and the number maxi of the maximum density parameter is checked.

(6) If $maxI = 1$, then the sample x_i that corresponds to the sample set U of S_i is the clustering center, and Step (10) is performed.

(7) If maxI≠1, then the sample x_i of the maximum density parameter S_i that corresponds to the sample set U in S is determined, and $i = 1, 2\ldots n$.

(8) $sum_i = \sum_{i=1}^{n} d(x_i, x_j)$ is solved, where $d(x_i, x_j) \le MeanDist$ and $j = 1,2\ldots n$, and they will be placed in set SUM.

(9) $sum_i = \min(SUM)$, and the sample in set U is selected as the clustering center.

(10) All density parameters that correspond to the sample of $d(x_i, x_j) \le MeanDist$ are deleted from set S.

(11) Steps (4)–(10) are repeatedly executed until k clustering centers are selected.

4. Result Analysis and Discussion

The corpus of this thesis is the Sogou corpus with title, main body, and class, which is downloaded from http://www.sogou.com. This corpus includes nine classes, which are information technology (IT), health, education, finance, military, tourism, culture, sports, and entertainment. This thesis selects 30 texts for each class.

4.1 Analysis of Experimental Results

For the experiment, thirty texts are selected for each class of the above corpus, i.e., 1–30 are about IT classification numbers, 31–60 are health classification numbers, and the rest is selected in a similar way.

The dk-means and Adk-means algorithms are used to cluster the 270 texts. The text number of each class is given in Table 1, and the pre-clustering number $k = 9$. The clustering results of the dk-means and Adk-means algorithms are shown below.

Category	IT	Health	Education	Finance	Military
No.	1-30	31-60	61-90	91-120	121-150
Category	Tourism	Sports	Culture	Entertainment	
No.	151-180	181-210	211-240	241-270	

Table 1. Text classification number

The clustering results of the dk-means and Adk-means algorithms in Table 2 are compared with the accurate classification numbers in Table 1. The accuracy and integrity of the dk-means algorithm in actual text clustering are higher than those before improvement.

4.2 Evaluation of the Clustering Effect

To compare the clustering results in Table 2, F is used to measure the clustering effect, and results are shown in Table 3.

Two evaluation parameters—accuracy and recall rate—of the Adk-means algorithm are higher than those of the

Algorithm	Clustering result	Quantity
Dk-Means	IT No.: 1, 2, 5, 7-15, 17, 22-24, 25-27, 31, 34, 36, 38, 50, 57, 213-216, 230,261	32
	Health No.: 28, 29, 32, 33, 37, 39, 40, 41, 46-48, 41-56, 59, 63, 100, 130, 139, 141, 170, 187, 190, 197, 206, 229, 250, 252, 270	33
	Education No.: 16, 19, 35, 42, 43, 58, 60-62, 64-69, 75-77, 82-89, 91, 105, 116, 117, 120, 128, 131, 194, 200, 205, 208, 231, 248, 254, 266	42
	Finance No.: 30, 49, 70, 71, 73, 90, 96, 95, 97, 98, 99, 101, 103, 106, 107, 109-115, 119, 136, 175	25
	Military No.: 3, 4, 6, 118, 122-124, 126, 127, 129, 132, 134, 135, 137, 138, 140, 142,147, 149, 150, 161, 171, 195-199, 230, 261	33
	Tourism No.: 17, 18, 20, 21, 151-155, 157-160, 162-169, 173, 174, 176-180, 207	29
	Sports No.: 15, 74, 104, 125, 133, 172, 181-186, 188, 189, 191-193, 201-204, 207, 210, 251, 255, 258, 259, 260	28
	Culture No.: 78-81, 92-94, 102, 108, 211, 212, 217-228, 232-240, 253, 256, 257	32
	Entertainment No.: 70, 71, 73, 148, 209, 241-247, 262-265, 267-269	19
Adk-Means	IT No.: 1, 2, 4-8, 10-15, 17-30, 38, 89	29
	Health No.: 31-37, 39, 41-45, 47, 48, 50-55, 57-60, 152, 200	27
	Education No.: 9, 61-80, 82-88, 90, 133	30
	Finance No.: 81, 91-105, 107-120, 215, 234, 242	33
	Military No.: 16, 46, 121, 123-146, 148, 150, 153	30
	Tourism No.: 49, 56, 151, 154-180, 240	31
	Sports No.: 181-189, 191-199, 201-210	28
	Culture No.: 3, 40, 106, 122, 149, 190, 211-214, 216-233, 235-239	31
	Entertainment No.: 147, 237, 241, 243-270	31

Table 2. Clustering algorithm results

dk-means algorithm. The high accuracy rate indicates that the improved Adk-means algorithm is more accurate than that before the improvement. The recall rate indicates that the integrity of the Adk-means algorithm is higher than that before the improvement. The improved clustering algorithm not only improves its own clustering performance but also considers the weight of the same part of speech, and this improved algorithm enhances the performance of web text clustering. Hence, the improved Adk-means algorithm greatly advances the text clustering performance when clustering web text.

Other indexes that measure the clustering effect are compared. Figure 3 shows that the purity and mutual information values of the improved algorithm are obviously higher than those of the dk-means algorithm, thereby indicating that the accuracy, integrity, and performance of partition cluster of this algorithm are greatly improved. The entropy of the improved algorithm is lower than that of the original algorithm, which indicates that the polymerization of the Adk-means algorithm is higher with higher similarity within class. Overall, the performance of the improved algorithm is obviously higher than that of the original algorithm.

	Dk-means algorithm		Adk-means algorithm	
	Accuracy rate	Recall rate	Accuracy rate	Recall rate
IT	69.36%	63.33%	83.10%	80.78%
Health	72.14%	76.12%	92.59%	83.31%
Education	67.52%	77.01%	86.67%	92.33%
Finance	72.31%	73.54%	87.88%	86.67%
Military	71.01%	66.67%	89.23%	85.28%
Tourism	74.08%	72.00%	86.66%	86.62%
Sports	64.29%	60.01%	85.43%	82.54%
Culture	65.63%	70.32%	81.25%	87.31%
Entertainment	73.11%	64.67%	87.06%	82.00%

Table 3. F measurement value

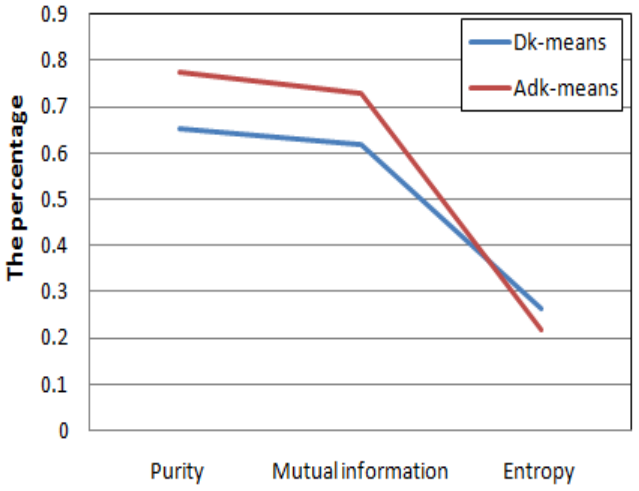

Figure 3. Experimental analysis

5. Conclusion

A large quantity of information on network is semi-structured or non-structured data in the advent of the Internet. Text information needs to be clustered to ensure convenient information search for users. This study proposes a k-means algorithm (Adk-means) based on average density optimization to improve the accuracy and integrity of web text clustering. Experimental results show that the Adk-means algorithm has a higher convergence rate, stronger stability, and higher clustering precision than the k-means algorithm based on density, and it eliminates the sensitivity of clustering results to isolated points. This algorithm is applied to a web text clustering model, and it completes the clustering and classification of web texts with high polymerization and similarity within class to improve the accuracy and integrity of web text clustering. The proposed algorithm can obtain target information rapidly and accurately. Thus, it is a highly efficient information retrieval technology. The main results

are as follows:

(1) A web text mining system model was constructed. The system model is divided into three parts. The first part is the text preprocessing module that is used to complete data preprocessing, word segmentation, selection of eigenvalues, and construction of eigenvectors. The second part is the text clustering module that obtains the clustering center. The third part is corpus classification.

(2) The basic definition, idea, and realization flow of the Adk-means were provided. Data on network are semi-structured or non-structured. Text information needs to be clustered to ensure convenient information search for users. The Adk-means algorithm is an efficient web text clustering method.

(3) Experimental verification. This study selected the data in the Sogou corpus to conduct an experimental test. Results showed that the Adk-means has high accuracy and integrity in actual text clustering, thereby verifying the effectiveness of this algorithm.

Further research will be conducted in the future. We aim to find better methods to further improve system efficiency and conduct intelligent web text mining application research to make progress for Internet applications.

Acknowledgements

The study was supported by Scientific Research Fund of Heilongjiang Provincial Education Department, China(No.12541058).

References

[1] Lee, Suiang-Shyan., Lin, Ja-Chen J. (2013). Fast K-means clustering using deletion by center displacement and norms product (CDNP). *Pattern Recognition and Image Analysis*, 23 (2), 199-206.

[2] Celebi, Emre M., Kingravi, Hassan, A., Vela, Patricio A.J. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40 (1), 200-210.

[3] Sharifi, Aboosaleh, M., Amirgholipour, J. (2015). Intrusion Detection Based on Joint of K-Means and KNN. *Journal of Convergence Information Technology*, 10 (5), 42-51.

[4] Jia, Lu, J. (2009). Research into K-means Clustering Algorithm Based on Dynamic Tunneling System.*Journal of Chongqing Normal University*, 26 (1), 73-77.

[5] Napoleon, D., Ganga Lakshmi, P.J(2010). An Enhanced k-means algorithm to improve the Efficiency Using Normal Distribution Data Points. *International Journal on Computer Science and Engineering*, 2 (7), 2409-2413.

[6] Zhiyong, Zhou D. (2007). *Study on Clustering Analysis Algorithm*. Master's thesis of Hebei University, Shijiazhuang, China.

[7] Changzheng, Xing.,Hao, Gu, J. (2014). K-means algorithm based on average density optimizing initial cluster centre. *Computer Engineering and Applications*, 50 (20), 135-138.

[8] Chen, Anhua., Pan, Yang. Jiang. Lingli, J. (2013). Improving K-means Clustering Method in Fault Diagnosis based on SOM Network. *Journal of Networks,* 8 (3), 680-687.

[9] Bharti, Kusum., Jain, Shweta., Shukla, Sanyam, J. (2010). Fuzzy K-mean Clustering Via Random Forest For Intrusiion Detection System. *International Journal on Computer Science and Engineering*, 2 (6), 2197-2200.

[10] Jun, Sunghae., Park, Sang-Sung., Jang, Dong-Sik J(2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. *Expert Systems with Applications*, 41 (7), 3204–3212.

[11] Vinh, La The., Lee, Sungyoung., Park, Young-Tack., d'Auriol, J, Brian J. (2012). A novel feature selection method based on normalized mutual information. *Applied Intelligence*, 37 (1), 100-120.

[12] Li, Haiyang., He, Hongzhou., Wen, YonggeJ(2015). Dynamic particle swarm optimization and K-means clustering algorithm for image segmentation. *Optik - International Journal for Light and Electron Optics*, 126 (24) , 4817–4822.

[13] Wang, Qinglei., Qian, Yanan., Song, J. Ruihua(2013). Mining subtopics from text fragments for a web query. *Information Retrieval*, 16 (4), 484-503.

[14] Kumar, Ajay., Kumar, Shishir., Saxena, Sakshi J(2012). An Efficient Approach for Incremental Association Rule Mining through Histogram Matching Technique. *International Journal of Information Retrieval Research* (IJIRR), 2 (2), 29-42.

[15] Cheng, Hong., Zhou, Yang., Huang, Xin J.(2012). Clustering large attributed information networks: an efficient incremental computing approach. *Data Mining and Knowledge Discovery*, 25 (3), 450-477.