ISSN No. 0976-5697

Volume 5, No. 5, May-June 2014

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Study of Recent Advancement in Document Clustering

Durgesh Nandan Dixit
Department of Computer Science
Madhav Institute of Technology & Science
Gwalior, India

R. K. Gupta
Department of Computer Science
Madhav Institute of Technology
Gwalior, India

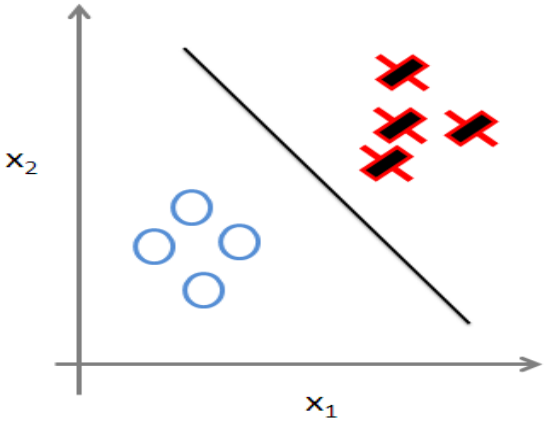
Abstract: Today data on internet is increasing at an exponential rate. Internet users are acting as a Data Producers and are pouring the internet with the lot of documents. Information retrieval (IR) is used to retrieve the more preferred information over the less preferred information. Thus Document clustering is a subset of the larger field of data clustering, which inherit concepts from the fields of information retrieval (IR) and machine learning (ML). In this paper we have analyzed the current state of document clustering research. A study of the algorithms is performed and directions for future research are also discussed.

Keywords: data mining; text mining; text clustering; k-means; non-negative factorization,PCA

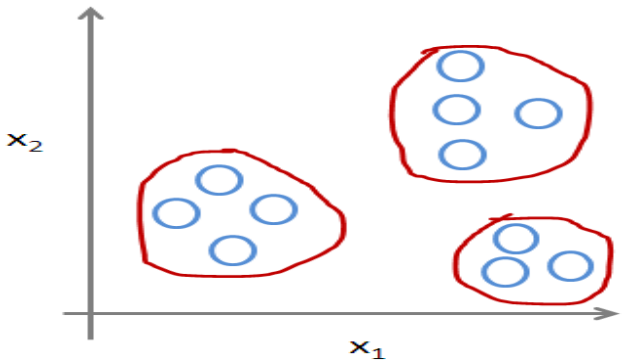
I. INTRODUCTION

With the rapid development of the Internet Technology, the electronic information is increasing day by day. It has become a tedious task for the users that how to retrieve the more preferred information over the less preferred information. Information Retrieval technology can help users to alleviate this problem. Document clustering is a subset of Information Retrieval technology which works as a filter to filter out the more useful information from less useful information [1]. It can serve as a fundamental and effective technique for efficient document management, document summary, navigation, and retrieval of large number of useful documents. Fast and high-quality document clustering algorithms play a vital role towards this goal via cluster-driven, dimensional reduction of vector space, term-weighting of document, or query expansion [2].

The process of document clustering aims to discover natural groupings, and present an overview of the categories in a collection of documents. In the field of machine learning, this is known as unsupervised learning. Clustering should not to be confused with classification. In a classification problem, the number of categories is known priori, and documents are assigned to these categories only. Classification is known as supervised learning. Conversely, in a clustering problem, the number of categories does not know in advance [3]. This distinction is illustrated in Figure 1.



(a) Classification



(b) Clustering

Figure 1: In (a), Two categories are known priori, and documents are assigned to each of them. In (b), an unknown number of groupings must be inferred from the data based on a similarity criterion

The document model is discussed in Section 2 which is a first challenge in a clustering problem is to determine which features of a document are to be considered discriminatory. In section 3 K-means algorithm for data clustering is described. Section 4 describes a number of matrix factorization techniques to reduce the number of the dimensions in term document matrix. Section 5 describes commonly used metrics. Finally, the conclusion is given in section 6.

II. VECTOR SPACE MODEL

In the vector model, a collection of n documents with m unique terms is represented as a m x n term-document matrix (where each document is a vector of m dimensions) [5].

A. Preprocessing:

Preprocessing take plain text document as an input and output a set of tokens to be included in the vector model. Preprocessing consists of following these steps:

a. Filtering:

The process of removing special characters, punctuation, helping verb, article, and conjunction words

that are not hold any discriminative power under the vector space model [6].

III. TOKENIZATION

The process of splitting the sentences into individual tokens, typically words. In this, to pick significant terms or chunks (sequences of words), such as noun phrases, the grammatical structure is to be parsed.

A. Stemming:

The process of reducing words to their base form, or stem. For example, the words “clustering”," clustered", “clusters" are all reduced to the stem “cluster."

B. Stopword Removal:

The process of removing words that do not convey any meaning, are removed from the text. The typical approach taken in removing stopwords is to compare each term with a compilation of known stopwords.

Another method is to first apply a part-of-speech tagger and then remove all terms that are not nouns, verbs, or adjectives.

C. Pruning:

The process of removing those words that appears with very low frequency throughout the corpus. These low frequency words do not contribute in separation of documents into clusters. Sometimes words which occur too frequently throughout the corpus are also rejected because these words exist in almost each and every cluster so does not hold any discriminating power [7].

D. An Example:

To illustrate this process, consider the first few sentences of the abstract of [8]:

In existing unsupervised methods, Latent Semantic Analysis (LSA) is used for sentence selection. However, the obtained results are less relevant, because singular vectors are used as the bases for sentence selection from given documents, and singular vector components can have negative values. Here we are using a new unsupervised method using Non-negative Matrix Factorization (NMF) to select sentences for automatic generic document summarization. This method uses positive and non-negative constraints, which are more similar to the human observation process. As a result, the method selects more relevant sentences for generic document summarization than those selected using LSA.:

Existing unsupervised method Latent Semantic Analysis use sentence selection However obtain result meaningful singular vector use base sentence selection give document singular vector component negative value propose unsupervised method us non negative matrix factor select sentence automate generic document summarization proposed method use positive and non negative constraint similar human cognition process result method select meaningful sentence generic document summary select us LSA.

The preprocessed abstract has many unique terms as the original, and can improve retrieval performance. designations.

IV. EXTENSION TO K-MEANS

K-means is very popular algorithm for document clustering. Let we select our initial centroids by randomly choosing *K* documents. Let $\{a_1, a_2, \ldots\ldots a_m \in$ is a training set of m documents [9]. We randomly initialize k cluster centroids $\mu_1, \mu_2 \ldots\ldots \mu_k \in$. Figure 2 shows four iteration of K-means algorithm. In each iteration, we assign each training example to the closest cluster centroid.

Repeat {

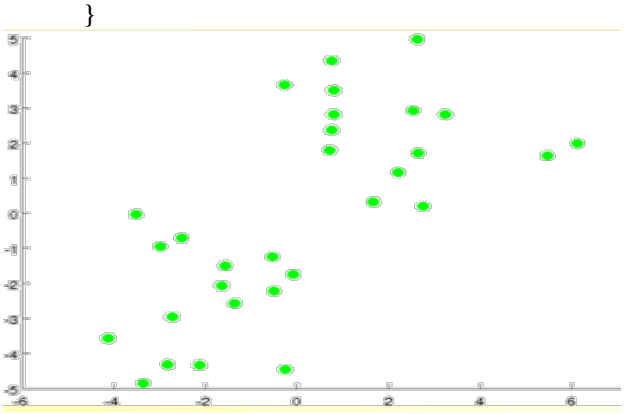
For i = 1 to m

Center$^{(i)}$: = index (from 1 to k) of cluster centroid closest to

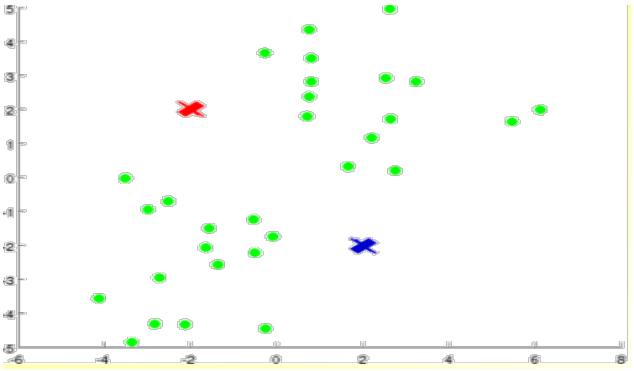
For k = 1 to K

μ_k: = average of points assigned to cluster k

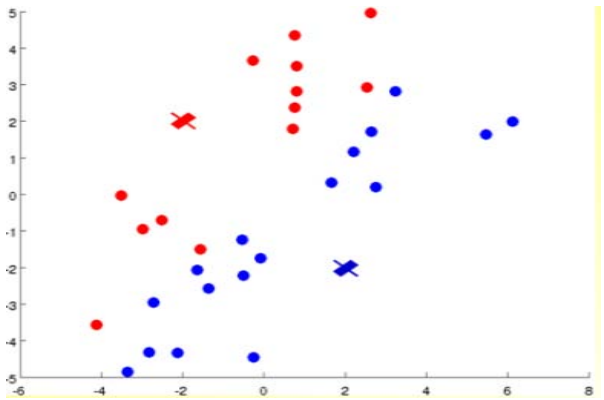
}



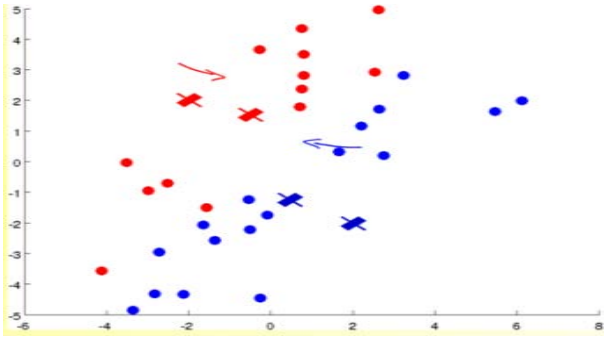
(a)



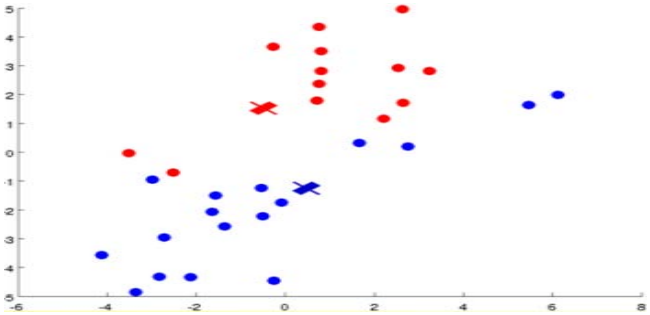
(b)



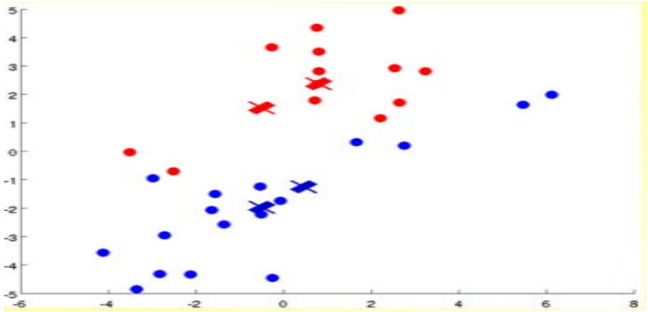
(c)



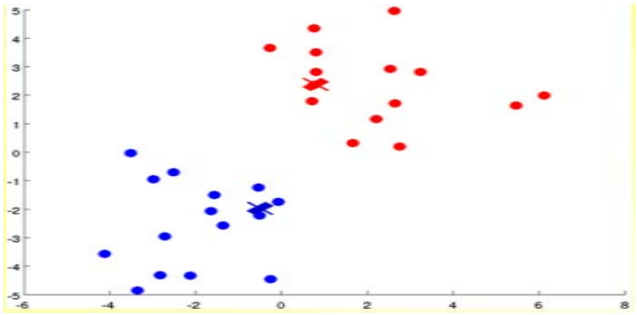
(d)



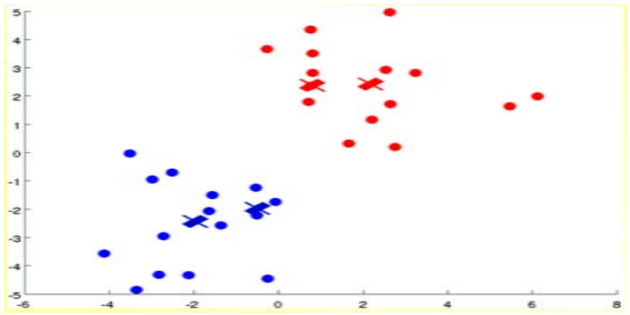
(e)



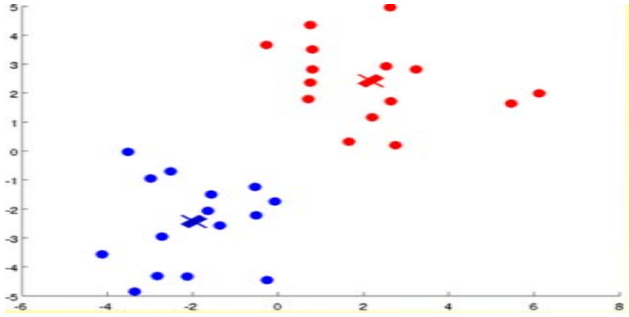
(f)



(g)



(h)



(i)

Figure 2: K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-i) Illustration of running four iterations of k-means.

K-means optimization objective-We have to minimize the cost function to increasing the performance of K-means algorithm.

$$Cost\ (center,\ \mu) = \frac{\sum_{i=1}^{m} \| a^{(i)} - \mu_{center(i)} \|^2}{m} \quad (1)$$

The cost function is a non-convex function, and so coordinate descent on this function is not guaranteed to converge to the global minimum [9].

A. Choosing the value of K:

As we increase the no of clusters K cost function decreases but up to some extent. As shown in figure 3 cost function remain unchanged after increasing the number of clusters more than three. So value of k should be choose very carefully as minimum as possible.

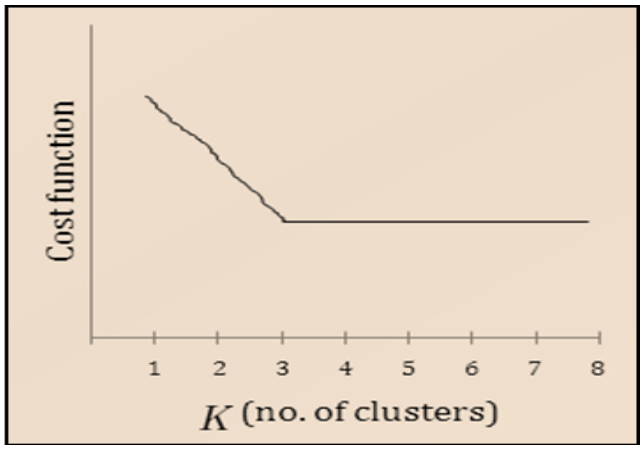


Figure 3

Problem with k-means is that it can be susceptible to local optima. One common thing to do is run k-means many times (using different random initial values for the cluster centroids μ). Then, out of all the different clustering found, pick the one that gives the lowest cost.

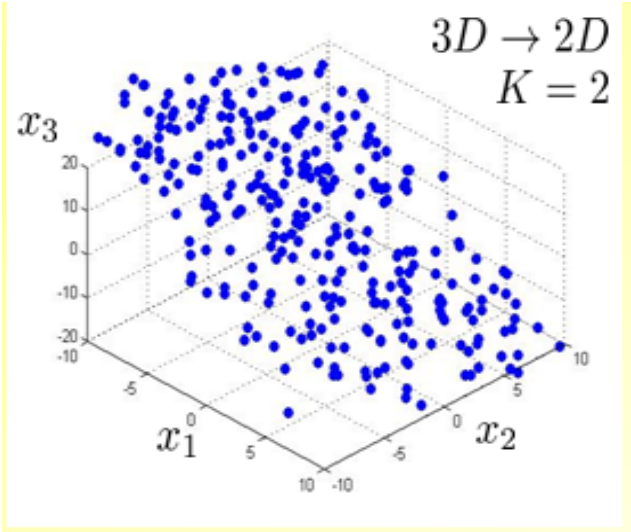
V. DIMENSIONALITY REDUCTIONS

With the help of preprocessing we can achieve significant reduction in the size of the vector space, although for higher efficiency more reduction is needed. This section describes two matrix factorization techniques that have been

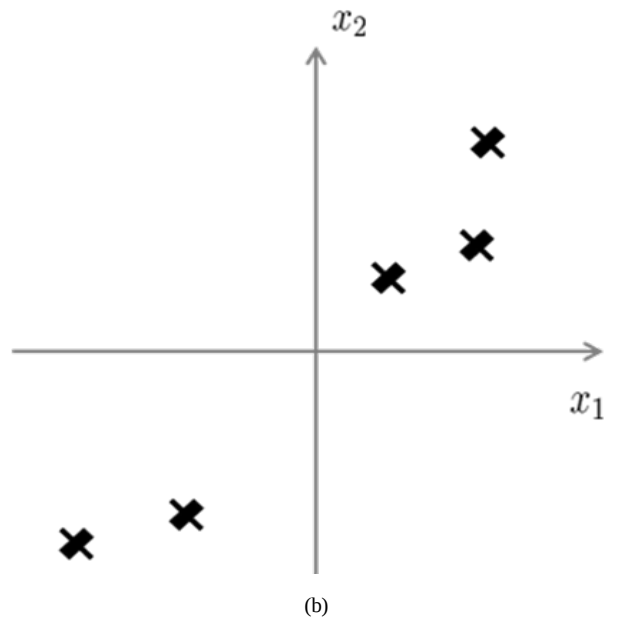
shown to not only significantly reduce the size of document vectors but also to increase clustering accuracy [10].

A. Principal Component Analysis:

Principal components are orthogonal projections that together explain the maximum amount of variation in a dataset. Let assumes a term-document matrix $A \in \Re^{m \times n}$. The goal of dimensionality reduction techniques is to produce a rank k approximation of A, A_k, while introducing controllable error.PCA seeks directions that represent data best in a $\Sigma|a_k - a|^2$. It reduces from n-dimension to k-dimension. Find k vectors $(\mu^{(1)}, \mu^{(2)}, \ldots\ldots\ldots\ldots \mu^{(k)})$ onto which to project the data, so as to minimize the projection error [11]. As shown in Figure 4 reduction from 3-dimension to 2-dimension



(a)



(b)

Figure 4: Reduce from 3-dimension to 2-dimension: Find a direction (a vector $\mu^{(1)}, \mu^{(2)} \in \Re^n$) onto which to project the data so as to minimize the projection error.

Principle components can be found by computing the singular value decomposition on the correlation matrix of the dataset. Correlation matrix can be computed as:

$$M = \frac{\sum_{i=1}^{n} (a^i) * (a^i)^T}{m} \quad (2)$$

Then Compute singular value decomposition of matrix M

$$M = U * S * V^T \quad (3)$$

Where S is a diagonal matrix and U and V are orthogonal matrices. Then take k dimensions of U matrix for clustering the documents.

$$a^i_{approx} = U^{'}_k * a^i \quad (4)$$

B. Choosing k (number of principal components):

Typically, choose k to be smallest value so that

$$\frac{\frac{1}{m} * \sum_{i=1}^{m} \| a^i - a^i_{approx} \|^2}{\frac{1}{m} \sum_{i=1}^{m} \| a^i \|^2} \leq 0.01 \quad (5)$$

It has been observed that PCA has two important properties that make it suitable for clustering: approximation and discriminate [11]. Approximation states that PCA introduces controllable error as dimensionality is reduced. The second property, discriminate, is more serious. Experiments have shown that PCA increases the ratio between similarity "*within the cluster*" and similarity between "*different clusters*". In other words, this indicates that in the reduced space clusters are more relevant. A problem with PCA is that the resulting approximation contains negative values. So for addressing this problem new dimensionality reduction method is introduced named as non negative matrix factorization.

C. Nonnegative Matrix Factorization:

Let assumes a term-document matrix $A \in \Re^{m \times n}$. In NMF we need to find two non-negative matrices P and Q that satisfy (6).

$$A \approx P^T * Q \quad (6)$$

Where Dimensions of P and Q are $m \times k$ and $k \times n$ respectively, where k is the reduced rank. Usually k is chosen to be much smaller than n, but more accurately, $0 < k << \min(m, n)$ and value of k also depend on application.

To achieve this we need to minimize cost function that measure the difference between A and PQ.

$$\min_{P,Q} \| A - PQ \|^2 \quad (7)$$

To minimize the cost function multiplicative update rule described in [12] as

a. Initialize P and Q with nonnegative values.
b. Iterate for each c, j, and i until convergence or after l iterations:

$$Q_{c,j} \leftarrow Q_{c,j} \frac{(P^T Q)_{c,j}}{(P^T P Q)^{+\epsilon}_{c,j}} \quad (8)$$

$$P_{i,c} \leftarrow P_{i,c} \frac{(PQ^T)_{i,c}}{(PQQ^T)_{i,c}^{+\in}} \quad (9)$$

Where small positive parameter equal to 10^{-9}, is added to avoid division by zero.

Random initialization of matrices P and Q creates problem with NMF is that the same data might produce different results on different runs [13] [14]

VI. EVALUATION METRIC

Two intuitive notions of performance (accuracy) are precision and recall. Recall is defined as the proportion of relevant documents that are retrieved out of all relevant documents available [1] [14],

$$\mathrm{Re}\,call = \frac{\mathrm{Re}\,levantDocument\,\mathrm{Re}\,trieved}{Total\,\mathrm{Re}\,levantDocument} \quad (10)$$

While precision is the proportion of retrieved and relevant documents out of all retrieved documents.

$$\mathrm{Pr}\,ecision = \frac{\mathrm{Re}\,levantDocument\,\mathrm{Re}\,trieved}{Total\,\mathrm{Re}\,trievedDocument} \quad (11)$$

Let R be recall and P be precision, then F-measure, which combines both recall and precision is defined as:

$$F\_measure = \frac{2*R*P}{R+P} \quad (12)$$

VII. CONCLUSION & FUTURE WORK

In this paper we have studied various techniques for text clustering algorithms. In future , we will perform various experiments using various standard datasets and will study the behavior of these algorithms under various constraints.

VIII.ACKNOWLEDGMENT

The authors specially wish to thank various research groups for making their dataset available for research purposes.

IX. REFERENCES

[1] D. Cutting, D. Karger, J. Pederson, and J. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of ACM SIGIR*, 1992.

[2] M.W. Berry, Z. Drmaˇc, and E. Jessup. Matrices, Vector Spaces, and Information Retrieval. SIAM Review, 41:335–362, 1999.

[3] Gerard Salton, "Developments in Automatic Text Retrieval," Science, 253:974-980, August 1991.

[4] Salton, G., Wang. A., & Yang, C.S., "A Vector Space Model for Information Retrieval," Communications ofthe ACM, 18(11): 613-620, November 1975.

[5] Tao Liu, Shenping Liu, Zheng Chen, & WeiYing Ma, "An Evaluation on Feature Selection for Text Clustering," Proc. of ICML-2003, Washington DC, 2003.

[6] Yang, Y., & Pedersen, J. O., "A Comparative Study on Feature Selection in Text Categorization," Proc. of ICML-97, pp. 412-420, 1997.

[7] Ju-Hong Lee, Sun Park, Chan-Min Ahn, Daeho Kim, Automatic generic document summarization based on non-negative matrix factorization, Information Processing and Management 45 (2009) 20–34.

[8] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

[9] Wyse, N., Dubes, R., & Jain, A.K., "A Critical Evaluation of Intrinsic Dimensionality Algorithms," Pattern Recognition in Practice, pp. 415-425, North-Holland, 1980

[10] Jolliffe, I.T., "Principal Component Analysis," Springer Series in Statistics, 1986.

[11] [11] D. Guillamet and J. Vitria. Determining a Suitable Metric when Using Non-Negative Matrix Factorization. In Sixteenth International Conference on Pattern Recognition (ICPR'02), Vol. 2, Quebec City, QC, Canada, 2002.

[12] W. Xu, X. Liu, and Y. Gong. Document-Clustering based on Non-Negative Matrix Factorization. In Proceedings of SIGIR'03, July 28-August 1, pages 267–273, Toronto, CA,2003.

[13] F. Shahnaz. Clustering Method Based on Nonnegative Matrix Factorization for Text Mining. Master's thesis, Department of Computer Science, University of Tennessee, Knoxville, TN, 2004.

[14] M.H. Dunham. Data Mining: Introductory and Advanced Topics. Prentice Hall, Upper Saddle River, NJ, 2003.