Dcoument Clustering in Research Literature Based on NMF and Testor Theory

Fang Li, Qunxiong Zhu
School of Computer Science and technology
Beijing University of Chemical Technology, BUCT
Beijing, China
e-mail: lifang@mail.buct.edu.cn

***Abstract*—The paper proposes a new way of comprising the NMF and testor theory to cluster research literature. NMF method is good at dealing with high dimensional documents and clustering, while testor theory is used to find the topic of each cluster. The whole process is described in detai through an example of ten abstracts of Chinese science literature from magazines relative to environmental science. In the end, a case study about automatic classification of a conference proceeding (in Chinese) is given. The result shows the effectiveness of the whole method.**

***Index Terms*—Document Clustering, NMF, Term-Document Matrix, Testor theory, Topic Discovery**

I. Introduction

There is a large and growing volume of published journals, conference proceedings and practitioner literature available for consideration and study in practically any academic subject. With such volumes of scholarly literature becoming increasingly and more easily available, it continues to be more important for the researcher to have the means at his disposal to intelligently and effectively sift through the piles of literature to identify and retrieve the gems of knowledge that pertain to his chosen subject. And at the same time it’s really exhausted to deal with so many documents manually for researchers. If computers can automatically organize a document corpus into a meaningful hierarchy, which enables an efficient browsing and navigation of the corpus, it will be great convenience to researchers and even all kinds of industry. Document clustering methods are good choices.

Document clustering methods can be mainly categorized into two types: document partitioning and agglomerative clustering. A lot of different algorithms have been proposed in the literature[7,8,9]. Hierarchical clustering is a very important concept for document clustering. Hierarchical clustering solutions have been primarily obtained using agglomerative algorithms, in which objects are initially assigned to its own cluster and then pairs of clusters are repeatedly merged until the whole tree is formed. However, partitional algorithms can also be used to obtain hierarchical clustering solutions via a sequence of repeated bisections. In recent years, various researchers have recognized that partitional clustering algorithms are well-suited for clustering large document datasets due to their relatively low computational requirements. And also, Some researchers have made comparative study [1] on those basic algorithms.

Although many kinds of methods have been extensively investigated for several decades, accurately clustering documents without domain-dependent background information, nor predefined document categories or a given list of topics is still a challenging task, because most of these algorithms do not really address problems like the very high dimensionality of the data, which requires the ability to deal with sparse data spaces or a method of dimensionality reduction. Non-negative matrix factorization (NMF)[2,12,13] solved this problem. NMF method is very fit for dealing with the high dimensionality reduction text data, and at the same time clustering them.

But a problem we should still face using NMF clustering method is the same with using other clustering technique, that is, we only cluster the data, but we can’t know what each cluster really mean. In order to determine at a glance whether the content of a cluster are of user interest or not, we import Testor Theory[3] to tag each clusters, and identifying distinct and finding representative topic of each cluster, and complete the process of topic discovery in research literature.

The rest of this paper is organized as follows. Section 2 briefly introduces the general approach of document representation in research literature. In section 3, we introduce the document clustering methods based on NMF. We described how to apply the Testor Theory for topic discovery in detail in section 4. And then a case study is deeply analyzed in section 5. In the end, section 6 summarizes the paper and outlines some interesting directions for future research.

II. Document representation in research literature

As we all know, research literature is a kind of unstructured document, and research literature is different from common documents in the following aspects:

- Generally the research literature has an abstract, which is short enough to analyze and also

Corresponding author: Qunxiong Zhu

doi:10.4304/jsw.6.1.78-82

represent the real meaning of the whole document.

- High dimensionality is still a problem for document representation.
- There are some special English texts mixed in Chinese research literature. So what we face are multilingual problems.

So, first we should select the English words mixed in the document, and store them into a database to manage. Second, we should deal with the rest part of the abstract.

Chinese text representation is different from English, because in case of Chinese document text, there is no obvious space between Chinese words. So the first step is word-segmentation. We use the word segmentation software (ICTCLAS), developed by Chinese Academy of Science, to make Chinese word segmentation. A text document is represented by a set of words, i.e. a text document is described based on the set of words contained in it (bag-of-words representation). In order to be able to define at least the importance of a word within a given document, usually a vector representation is used, where for each word a numerical ”importance” value is stored.

With the standard vector space model, a set of documents S can be expressed as a m×n matrix V, where m is the number of terms in the dictionary and n is the number of documents in S. Each column Vj of V is an encoding of a document in S and each entry Vij of vector Vj is the significance of term i with respect to the semantics of Vj, where i ranges across the terms in the dictionary.

In order to illustrate the methods conveniently, we select a simple document set as our data corpora: 10 abstracts of Chinese science literature from magazines relative to environmental science. These literatures are about soil, contamination processing and biological science. We don’t consider the term weight for convenience.

We extract 20 keywords from the 10 abstract documents manually to construct the term-document matrix M(Table 1).

III. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization (NMF) is a matrix factorization technique to realize the dimension reduction, and at the same time, the cluster membership of each document can be easily identified from NMF. Even though, if we apply a traditional data clustering methods after NMF processing. With the standard vector space model, a set of documents S can be expressed as a m×n matrix V, where m is the number of terms in the dictionary and n is the number of documents in S. Each column V_j of V is an encoding of a document in S and each entry V_{ij} of vector V_j is the significance of term i with respect to the semantics of V_j, where i ranges across the terms in the dictionary. The NMF problem is defined as finding a low rank approximation of V in terms of some metric (e.g., the norm) by factoring V into the product (WH) of two reduced-dimensional matrices W and H. Each column of W is a basis vector, i.e., it contains an encoding of a semantic space or concept from V and each column of H contains an encoding of the linear combination of the basis vectors that approximates the corresponding column of V. Dimensions of W and H are m×k and k×n respectively, where k is the reduced rank or selected number of topics. Usually k is chosen to be much smaller than n, but more accurately, k<<min(m, n). Finding the appropriate value of k depends on the application and is also influenced by the nature of the collection itself.

TABLE I. ORIGINAL TERM-DOCUMENT MATRIX X

Term	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Soil(土壤)	1	1	2	0	0	0	0	0	0	0
Ground(土地)	0	1	0	0	0	0	0	0	0	0
Environment (环境)	0	1	0	0	0	0	0	0	0	0
Basification (碱化)	0	0	1	0	0	0	0	0	0	0
Chemistry(化学)	0	0	1	0	1	0	0	0	0	0
Effluvium(恶臭)	0	0	0	1	0	0	0	0	0	0
Contamination (污染物)	0	0	0	1	1	0	0	0	0	0
Biology(生物)	0	0	0	0	1	0	0	0	0	0
Craftwork(工艺)	0	0	0	0	1	0	0	0	0	0
Sewage(污水)	0	0	0	0	0	1	1	0	0	0
Dispose(处理)	0	0	0	0	0	1	0	0	0	0
Filter(过滤)	0	0	0	0	0	0	1	0	0	0
Feasibility (可行性)	0	0	0	0	0	0	1	0	0	0
Gardens(园林)	0	0	0	0	0	0	0	1	0	0
Greenbelt(绿地)	0	0	0	0	0	0	0	1	0	0
Plant(植物)	0	0	0	0	0	0	0	1	0	1
Research(研究)	0	0	0	0	0	0	0	1	1	0
Country(山地)	0	0	0	0	0	0	0	0	1	0
Zoology(生态)	0	0	0	0	0	0	0	0	1	1
Virescence(绿化)	0	0	0	0	0	0	0	0	0	1

Common approaches to NMF obtain an approximation of V by computing a (W, H) pair to minimize the Frobenius norm of the difference V-WH. The problem can be cast in the following way—let $V \in R_{m\times n}$ be a nonnegative matrix and $W \in R_{m\times k}$ and $H \in R_{k\times n}$ for 0< k<<min(m, n). Then, the objective function or minimization problem can be stated as:

$$\min_{W,H} \| V - WH \|_F^2 \qquad (1)$$

with $W_{ij} > 0$ and $H_{ij} > 0$ for each i and j.

The matrices W and H are not unique. Usually H is initialized to zero and W to a randomly generated matrix where each $W_{ij} > 0$ and these initial estimates are improved or updated with alternating iterations of the algorithm.

The iterative algorithm shown in formula (2), (3) can get a local optimal result.

$$H_{cj} \leftarrow H_{cj} \frac{(W^T V)_{cj}}{(W^T WH)_{cj} + \varepsilon} \qquad (2)$$

$$W_{ic} \leftarrow W_{ic} \frac{(VH^{T})_{ic}}{(WHH^{T})_{ic} + \varepsilon} \qquad (3)$$

Applying NMF on the matrix M in table 1, we get $W_{20\times3}$ (basis matrix)and $H_{3\times10}$ (coefficient matrix) matrix. For the convenience of analysis, the coefficient matrix H is shown in Table 2.In the new semantic space formed by matrix W, the document can be viewed as a combination of basis vectors. Thus, NMF can be used to organize text collections into partitioned structures or clusters directly derived from the non-negative factors. The cluster sets are: Cluster1={D1,D2,D3},Cluster 2={D4,D5,D6,D7} and Cluster 3={D8,D9,D10}.

TABLE II. TABLE 3 COEFFICIENT MATRIX H FOR THE EXAMPLE OBTAINED BY NMF

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
1.57	2.13	4.17	0.00	0.14	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.02	4.02	2.94
0.00	0.00	0.30	2.03	3.27	0.16	0.26	0.00	0.00	0.00

Then We could also do a general clustering based on the H matrix using Pearson's coefficient. After that, those documents with higher relativity are grouped together, whereas those with lower relativity are separated. The cluster sets are:

Cluster1={D8,D9,D10},Cluster2={D1,D2,D3},Cluster3={D4,D7,D6,D5}, as presented in Figure 1.

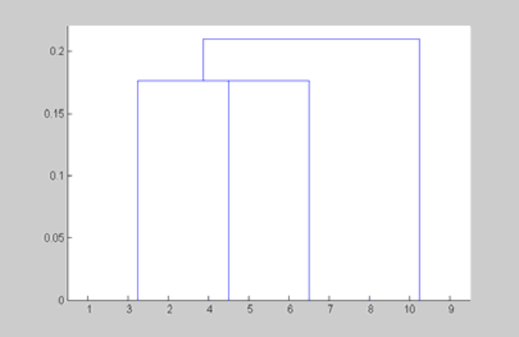

Figure 1.Clustering Result with Pearson's coefficient after NMF

IV. TOPIC DISCOVERY WITH TESTOR THEORY

Testor theory is used to give each cluster a tag after clusters are formed. The Testor Theory had its beginnings in the middle of the fifties, when the concept of testor appeared in the works of Yablonskii and Cheguis[3] for the first time. The aim of these works was the development of mathematical logical methods to find faults in electrical circuits. Their practical interest was given by the fast development accomplished in the field of Electronic Computing in these years. To find a fault in complicated logical circuits is a very hard task; hence, the natural idea of attempting to simplify this. Then, in the field of practical applications, some work about testor theory appeared, including the concept of testor for contact outlines Their applicability soon transcended and testor began to be applied as a useful tool in the solution of pattern recognition problems[10,11]. Here, we use testor theory to select the most key words in each cluster [3][4].

Let us denote by U the universe of objects in study, each of them described in terms of features $M = \{x_1,...,x_n\}$ and grouping the classes $K_1,...,K_n, r > 2$, which are not necessarily disjoint. The feature sets $\tau=\{x_{i1},...,x_{im}\}$ of matrix M is called a testor, if after deleting from M all columns except $\{i_1,...,i_m\}$, all rows of M corresponding to distinct classes are different. A testor is called typical if none of its proper subsets is a testor.

The basic idea is the following: A testor is a feature subset, which does not confuse any pair of subdescription from different classes. Moving, from a testor to a typical testor(elimination features, when it is possible) we get an irreducible combination of features, where each feature is essential in order to keep differences between classes. This property distinguishes each typical testor. It is nature to suppose that if a feature many times in different typical testors, it is more difficult to disregard it. That is we could say it is more useful to differentiate between classes.

Then we have the following steps.

A. To Construct Learning Matrix

For each cluster c, we construct a learning matrix LM(c), whose columns are the most frequent terms in the representative $\bar{c}$, and its rows are the representatives of all clusters, described in terms of these columns. In order to calculate the typical testors, we considered two classes in the matrix LM(c). The first class is only formed by $\bar{c}$ and the second one is formed by the other cluster representatives. Notice that our goal is to distinguish the cluster c from the other clusters. The LM(c) matrix is as shown in table 3：

TABLE III. LEARNING MATRIX LM(C) OF CLUSTER C

	Soil (土壤)	**Ground (土地)**	**Environment (环境)**	**Basification (碱化)**	**Chemistry (化学)**
$\bar{c}_1$	0.4	0.1	0.1	0.1	0.1
$\bar{c}_2$	0	0	0	0	0.1
$\bar{c}_3$	0	0	0	0	0.1

B. To ConstructComparison Matrix

For the calculus of the typical testors of a matrix M, the key concept is to construct the comparison matrix. Comparison matrix could be a matrix of similarity or a matrix of dissimilarity depending on the type of comparison criteria that are applied for each feature. In our case, the features that describe the documents are the terms and its values are the frequency of terms. The comparison criterion applied to all the features is

$$d(v_{i_k}, v_{j_k}) = \begin{cases} 1 & if\, v_{i_k} - v_{j_k} \geq \delta \\ 0 & otherwise \end{cases} \qquad (4)$$

Where v_{i_k} , v_{j_k} are the frequencies in the cluster representative i and j in the column corresponding to the term t_k respectively, and d is a user-defined parameter. As it can be noticed, this criterion considers the two values (frequencies of the term t_k) different if the term t_k is frequent in cluster i and not frequent in cluster j. The comparison matrix is shown in table 4.

TABLE IV. COMPARISON MATRIX

	Soil (土壤)	**Ground (土地)**	**Environment (环境)**	**Basification (碱化)**	**Chemistry (化学)**
$\bar{c}_1,\bar{c}_2$	1	0	0	0	0
$\bar{c}_1,\bar{c}_3$	1	0	0	0	0

C. Cluster Representation

From table 3, we can find the typical testors of this matrix are:{soil(土壤)}, so the key term soil(土壤) are the most representative word in cluster1, which we can use to tag the cluster. In the same way, we can find another two key terms contamination (污染物) and Zoology (生态) representing cluster 2 and cluster 3.

V. CASE STUDY AND RESULT ANALYSIS

As a test of using the former methods to analyze existing literature, we decided to use the article abstracts published in the proceeding of IEEE ICC 2008 China Forum. So we can compare the clustering result obtained from our methods with the category classified manually, and also compare the label words with the known topic name.

There are 107 papers in the proceeding, which have been classified into five categories manually as shown in Tab.5.

TABLE V. CATEGORY OF PROCEEDING OBTAINED MANUALLY

C1	C2	C3	C4	C5
Wide band (宽带技术)	Wireless (无线技术)	Transport of digital video (数字视频传输)	Increment Service (增值服务)	Other (其他)
19	32	9	46	1

For this data collection, after preprocessing, word-segmentation, clustering with NMF and Pearson's coefficient, and labeling with testor theory, we get seven categories as shown in tab.6.

TABLE VI. CATEGORY OF PROCEEDING OBTAINED BY NMF+ TESTOR METHOD

T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Wide band (宽带)	Physical layer, high layer (物理层，高层技术)	Wireless (无线技术)	digital video (数字视频)	Multicast transport (组播传输)	Increment Service (增值服务)	Other (其他)
24	15	25	10	5	50	5

As we all know, the physical layer technique and high layer technique are branches of the wireless communication, so the T2 and T3 should be the same class, and multicast transport are also the problem of digital video, so the categories are almost the same.

The accuracy of the clustering is assessed using the metric AC[5] defined by

$$AC = \sum_{i=1}^{n} \delta(d_i)/n \qquad (5)$$

Where $\delta(d_i)$ is set to 1 if d_i has the correct topic label after using testor theory, and set to 0 otherwise, and n is the number of documents in the collection. We wrote a small program in C to calculate the AC value of each category. The average value is 0.8324.

On the other hand, the topic name of each category is almost satisfying. We find the most representative terms such as wide band, wireless communication, and digital video, etc.

VI. SUMMARY AND FUTURE WORK

In text mining field, using NMF and testor theory altogether to deal with documents is a new idea. NMF method efficiently reduces the dimensionality of the high dimensional text, and at the same time, clusters them, while testor theory solves the problem of clustering label well. In the end we apply the whole method to a conference proceedings' classification, the result comparing to the category obtained manually shows that the method is effective.

Of course, during the process of clustering by NMF, we got only the first level of topic. In fact, we could get more detailed sub-topics in the same way. So how to design some effective algorithms to get more specific topic level are sill under study and will be the further job for us.

Besides, with the development and data mining and machine learning technique, how to apply some typical data mining and machine learning methods into topic discovery to help solve above problems is also an interesting and meaningful job.

ACKNOWLEDGMENT

Financial supports from National Natural Science Foundation of China (No: 60774079) and National High Technology Research and Development Program of China (863 Program) (No: 2006AA04Z184) are highly appreciated.

REFERENCES

[1] Steinbach M., Karypis G., Kumar V.(2000), “A Comparison of Document Clustering Techniques,” Proceeding of TextMining Workshop, KDD 2000.

[2] Lee D D, Seung H S, “Algorithms for non-negative matrix factorization,” Advances in Neural Information Processing Systems, 2001(13), pp. 556–562.

[3] Manuel Lazo-Cortes, Jose Ruiz-Shulcloper, Eduardo Alba-Cabrera, “An overview of the evolution of the concept of testor,” Pattern recognition, 2001(34), pp. 753-762.

[4] Aurora Pons-Porrata a, Rafael Berlanga-Llavori b, Jose Ruiz-Shulcloper c, “Topic discovery based on text mining techniques,” Information Processing and Management 2007(43), pp. 752 – 768.

[5] Xu, W., Liu, X., & Gong, Y, “Document-clustering based on non-negative matrix factorization,” In Proceedings of SIGIR_03, Toronto, CA, July 28 – August 1, 2003, pp. 267 – 273.

[6] X. Liu and Y. Gong. “Document clustering with cluster refinement and model selection capabilities”. Proceedings of ACM SIGIR 2002, Tampere, Finland, Aug. 2002.

[7] Y. Li, S.M. Chung, Parallel bisecting k-means with prediction clustering algorithm, The Journal of Supercomputing, 2007,39 (1) pp.19 – 37.

[8] Y. Li, S.M. Chung, J.D. Holt, Text document clustering based on frequent word meaning sequences, Data and Knowledge Engineering 2008, 64(1) pp.381 – 404.

[9] W. Song, S.C. Park, “A novel document clustering model based on latent semantic analysis”. Proceedings. of the Third Int’l Conf. on Semantics, Knowledge and Grid, 2007, pp. 539 – 542.

[10] Santiesteban, Y., Pons-Porrata, A.: LEX: a new algorithm for the calculus of typical testors. Mathematics Sciences Journal 2003, 21(1),pp. 85–95

[11] Aurora Pons-Porrata1, Reynaldo Gil-García1, and Rafael Berlanga-Llavori, “Using Typical Testors for Feature Selection in Text Categorization”. L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, 2007, pp. 643–652.

[12] A. Pascual-Montano, P. Carmona-Saez, M. Chagoyen, F. Tirado,J. Carazo, R. Pascual-Marqui, bioNMF: a versatile tool for nonnegative matrix factorization in biology, BMC Bioinformatics. 2006,7. pp.366.

[13] M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, R.J. Plemmons, Algorithms and applications for approximate nonnegative matrix factorization, Comput. Stat. Data Anal. 2007, 52 (1). pp: 155–173.

Fang Li was born in AnGuo, HeBei Province, China, on August 11, 1977. She received his B.S degree and Master degree in Computer Science and Technology from Beijing University of Chemical Technology in 1999 and 2002 respectively.

Currently, she is a Ph.D research candidate with Control theory and Engineering at Beijing University and Technology since 2006. Her research interest is data mining, information processing. She is also a teacher of BUCT currently. She has been engaged in data mining research for approximately ten years.

Qunxiong Zhu was born in WuXi, JiangSu Province, China, in 1960. Currently, he is a professor, Ph.D. supervisor of BUCT, and teaches and conducts research in the areas of Intelligence engineering, information processing, etc.