Ping-pong Document Clustering using NMF and Linkage-Based Refinement

Hiroyuki Shinnou, Minoru Sasaki

Ibaraki University,
4-12-1 Nakanarusawa, Hitachi, Ibaraki, Japan 316-8511
{shinnou, msasaki}@mx.ibaraki.ac.jp

Abstract

This paper proposes a ping-pong document clustering method using NMF and the linkage based refinement alternately, in order to improve the clustering result of NMF. The use of NMF in the ping-pong strategy can be expected effective for document clustering. However, NMF in the ping-pong strategy often worsens performance because NMF often fails to improve the clustering result given as the initial values. Our method handles this problem with the stop condition of the ping-pong process. In the experiment, we compared our method with the k-means and NMF by using 16 document data sets. Our method improved the clustering result of NMF significantly.

1. Introduction

Document clustering is the task of dividing a document's data set into groups based on document similarity. This is the basic intelligent procedure, and is important in text mining systems (Michael W. Berry, 2003). As the specific application, relevant feedback in IR, where retrieved documents are clustered, is actively researched (Hearst and Pedersen, 1996)(Kummamuru et al., 2004).

Non-negative Matrix Factorization (NMF) is a clustering method based on the dimensional reduction method, and is effective for the document clustering, in which a vector is high-dimensional and sparse. In this paper, we propose the ping-pong clustering method that NMF and the linkage based refinement are conducted alternately, in order to improve the initial clustering result generated by NMF.

The ping-pong clustering consists of two clustering methods to improve the given clustering result, and uses these two methods alternately to improve the clustering result step by step. The term "ping-pong clustering" is not used generally, but in the paper (Dhillon et al., 2002), this method was called by the "ping-pong strategy." So in this paper, we name this method as "ping-pong clustering." Each method in the ping-pong clustering can be used as a clustering method by itself. The ping-pong clustering produces a better result than the single clustering method.

The "local search" proposed by Dhillon is representative of the ping-pong clustering (Dhillon et al., 2002). That method combines the k-means and the "first-variation" to improve the clustering result. Ding showed that NMF and pLSI use the same object function, but their search methods are different. Thus, he proposed the ping-pong clustering to use them alternately (Ding et al., 2006). In this paper, we use NMF and the linkage based refinement for the ping-pong clustering. In this paper, we will refer to the linkage based refinement as "LBR" for short.

NMF is a dimensional reduction method(Xu et al., 2003). Let X to be the $m \times n$ term-document matrix, consisting of m rows (terms) and n columns (documents). If the number of clusters is k, NMF decomposes X to the matrix U and V^T as follow:

$$X = UV^T$$

where U is $m \times k$, V is $n \times k$ and V^T is the transposed matrix of V. And the matrix U and V are non-negative. In NMF, each k dimensional column vector in V is correspoing to the document. An actual clustering is usually conducted by using these reduced vectors. However, NMF does not need that clustering procedure. The reduced vector expresses its cluster because each column axis of V represents a topic of the cluster.

The matrix V and U can be obtained by using a simple iterative procedure with the initial matrix V_0 and U_0 (Lee and Seung, 2000). The initial matrix V_0 is corresponding to a clustering result. Thus, NMF can be regarded as the method to improve the given clustering result. That is, we can use NMF as a constitutive method of the ping-pong clustering. For document clustering, the ping-pong clustering using NMF hold great promise because NMF is effective for document clustering.

LBR is the method to refine the clustering result. It was proposed in the paper (Ding et al., 2001) in order to refine the clustering result produced by the spectral clustering method, Mcut. LBR defines an object function to measure the refinement degree in the case that data u in the cluster A moves to the cluster B. By using that object function, each data is reassigned to a cluster. LBR does not guarantee to improve the value of the object function used in clustering, but is actually effective to refine the clustering result produced by the spectral clustering method (Ding et al., 2001). It should be considered that LBR is also effective for the any clustering result. So we use LBR as another constitutive method of the ping-pong clustering.

A novelty of this research is the use of NMF in the ping-pong clustering. As previously mentioned, the ping-pong clustering using NMF holds great promise. However, the ping-pong clustering using NMF has often negative effects because NMF does not always improve the given clustering result. To overcome this problem, we devise the stop condition of the ping-pong. Concretely speaking, we judge whether the ping-pong stops or not, through the value of an object function of the clustering result produced by LBR. If the value is improved, we keep the ping-pong. Otherwise we stop the ping-pong, and output the clustering result that LBR produced in the previous application.

In the experiment, we compared our method with the k-means and NMF using 16 document data sets. We eval-

uated clustering results by entropy, and showed that our method is effective.

2. NMF

NMF decomposes the $m \times n$ term-document matrix X to the $m \times k$ matrix U and the transposed matrix of the $n \times k$ matrix V (Xu et al., 2003), where k is the number of clusters:

$$X = UV^T.$$

NMF attempts to find the k axes corresponding to the topic of the cluster, and represents the document vector and the term vector as a linear combination of found k axes. That is, the coefficient of the axis means the degree of relevance to the topic. After all, the matrix V represents the clustering result. Concretely speaking, The i-th document d_i is corresponding to the i-th row vector of V, that is

$$(v_{i1}, v_{i2}, \cdots, v_{ik}).$$

The cluster number is obtained from

$$\arg\max_{j \in 1:k} v_{ij}.$$

For the given term-document matrix X, we can obtain U and V by the following iteration (Lee and Seung, 2000).

$$u_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^TV)_{ij}} \tag{1}$$

$$v_{ij} \leftarrow v_{ij} \frac{(X^TU)_{ij}}{(VU^TU)_{ij}} \tag{2}$$

The u_{ij}, v_{ij} and $(X)_{ij}$ mean the i-th row and the j-th column element of U, V and X respectively.
After each iteration, U must be normalized as follow:

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}}. \tag{3}$$

The iteration stops by the fixed maximum iteration number, or the distance J between X and UV^T:

$$J = ||X - UV^T||_F \tag{4}$$

where $||\cdot||_F$ means the Frobenius norm. The Frobenius norm of the $m \times n$ matrix A is defined by

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} {a_{ij}}^2}.$$

Generally, the initial the matrix V_0 and U_0 are constructed by random values. However, the iteration of Eq.1 and Eq.2 converges only to a local optimum solution. So the final V and U vary by the initial values. As the result, the clustering accuracy depends on V_0 and U_0.
On the other hand, the matrix V_0 is corresponding to a clustering result, so NMF can be regarded as the method to improve the given clustering result. Therefore, by giving the better initial values, we can expect to get the better result through NMF.

3. LBR

We use LBR as another constitutive method of the ping-pong clustering. LBR is developed to refine the clustering result produced by the spectral clustering method, Mcut (Ding et al., 2001). The spectral clustering method suffers from the "skewed cut" problem. LBR is the countermeasure for that problem.
In this section, first, we briefly explain Mcut, and then LBR.
In Mcut, the data set is represented as a graph. Each instance data is represented as the vertex in the graph. If the similarity between the data A and B is not zero, the edge between A and B is drawn, and given the similarity as the weight of the edge. From the view of this graph, clustering is corresponding to the segmentation of the graph into some subgraphs by cutting edges. This cut is preferable such that the sum of weights of inside edges of the subgraph is large, and the sum of weights of cut edges is small. To find the ideal cut, the object function is used.
We define the similarity $cut(A, B)$ between the subgraph A and B as follow:

$$cut(A, B) = W(A, B). \tag{5}$$

The function $W(A, B)$ means the sum of weights of edges between A and B. And we define that $W(A) = W(A, A)$. The object function of Mcut is the following:

$$Mcut = \frac{cut(A, B)}{W(A)} + \frac{cut(A, B)}{W(B)} \tag{6}$$

The clustering task is to find A and B to minimize the above equation. This minimization problem can approximately be solved by solving an eigenvalue problem. The "skewed cut" problem occurs in finding this approximate solution.
Note that the spectral clustering method divides data set into two groups. If the number of clusters is larger than 2, the above procedure is iterated recursively.
LBR defines an object function to measure the refinement degree in the case that data u in the cluster A moves to the the cluster B. If the degree is positive, data u is moved to the cluster B. That object function is defined as follows:

$$\Delta l_{AB}(u) = l(u, B) - l(u, A),$$

where

$$l(u, X) = \frac{1}{|X|} \sum_{v \in X} sim(u, v).$$

The $sim(u, v)$ means the similarity between u and v. In the case of $\Delta l_{AB}(u) < 0$, the data u stays in the cluster A.
LBR is basically for the dual partitioning. Mcut iterates recursively the dual partitioning. Thus, after each iteration, LBR is conducted.
Next we explain the general LBR for the cluster number is $k(\geq 2)$.
The object function of Mcut for the clustering result $\{G_1, G_2, \cdots, G_k\}$ is as follows:

$$Mcut_K = \frac{cut(G_1, \bar{G_1})}{W(G_1)} + \frac{cut(G_2, \bar{G_2})}{W(G_2)} + \cdots + \frac{cut(G_k, \bar{G_k})}{W(G_k)} \tag{7}$$

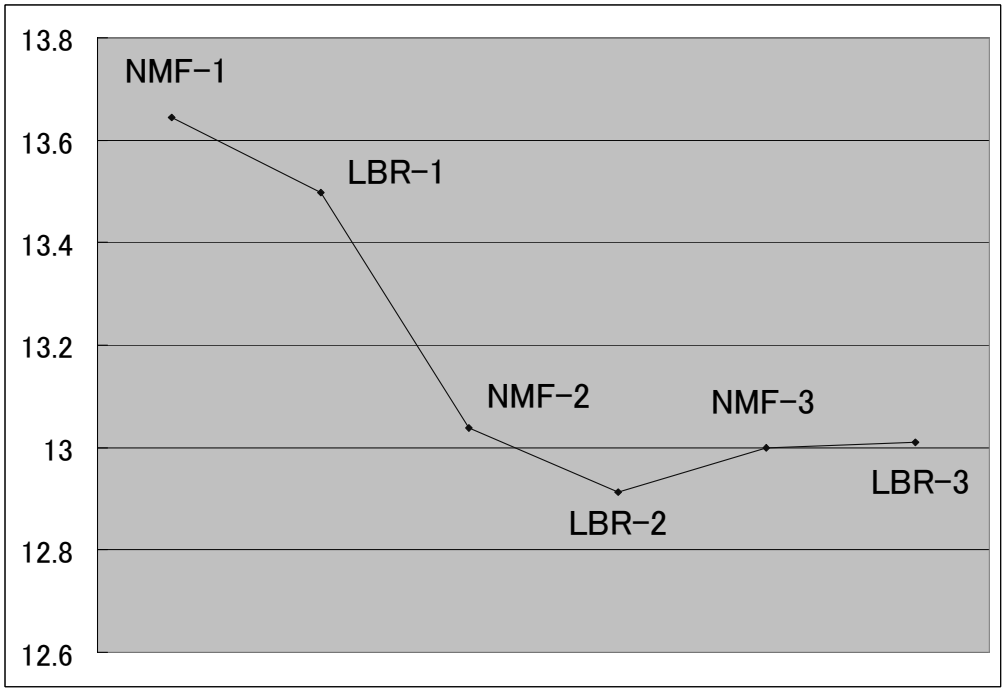

Figure 1: Value of the object function in the ping-pong clustering (1)

where the $\bar{G_k}$ means the complement of G_k. The smaller $Mcut_K$ is, the better it is.
Suppose the data u is a member of the cluster G_i. The $\Delta l_{ij}(u)$ is defined as follows;

$$\Delta l_{ij}(u) = l(u, G_j) - l(u, G_i).$$

Now we define the $\hat{i}$ as follows:

$$\hat{i} = \arg\max_j \Delta l_{ij}(u).$$

In the case of $i \neq \hat{i}$, the data u is moved from the cluster G_i to the cluster $G_{\hat{i}}$.
After conducting the above procedure for all data, we get the new clustering result $\{G_1, G_2, \cdots, G_k\}$. For this new clustering result, we iterate the above procedure. This iteration is stopped when the movement does not occur.
Note that LBR can not always improve the value of the object function Eq.7. That is, LBR is a heuristic method to improve the clustering result.

4. Ping-pong clustering

Our ping-pong clustering first conducts NMF, and get the clustering result. And then, the clustering result is improved by LBR. Using the improved clustering result, the initial matrix V_0 and U_0 of NMF are constructed as follows. If the cluster number of the i-th data is clustered into the c-th cluster in the improved clustering result. the i-th row vector of the V_0 is constructed as follow:

$$v_{ij} = \begin{cases} 1.0 & (j = c) \\ 0.1 & (j \neq c). \end{cases}$$

U_0 is constructed by XV_0. Using above V_0 and U_0 as initial matrices, NMF is conducted. As such ways, our ping-pong clustering conducts NMF and LBR alternately.
It is ideal that both of NMF and LBR can improve the given clustering result, but it is not guaranteed. Especially NMF often fails to improve the clustering result. So it is hard to use NMF in the ping-pong clustering.
To overcome this problem, we devise the stop condition of the ping-pong. Concretely speaking, we evaluate the value of the object function Eq.7 for the clustering result produced by LBR. If that value is improved, we keep the ping-pong process. Otherwise, we stop the ping-pong process, and output the clustering result produced by the previous LBR.
We show an example. The Figure 1 shows the result of our ping-pong clustering for the data set 'tr12' used in our experiment described in the next section. The vertical axis means the value of the object function (Eq.7).
First we conduct NMF, and obtain the clustering result (NMF-1). The value of the object function of NMF-1 is shown as 'NMF-1' in Figure 1. Next we conduct LBR by giving NMF-1, and obtain the clustering result (LBR-1). The value of the object function of LBR-1 is shown as 'LBR-1' in Figure 1. Next by using LBR-1, we construct the initial matrices V_0 and U_0. Next we conduct NMF by using V_0 and U_0, and obtain the clustering result (NMF-2). By iterating the above procedure, we obtain the clustering result (LBR-2). We compare values of the object function of LBR-1 and LBR-2. In this case, LBR-2 is smaller, so we keep the ping-pong, and obtain the clustering result (LBR-3). We compare values of the object function of LBR-2 and LBR-3. Now LBR-3 is larger than LBR-2, so we stop the ping-pong, and output the clustering result (LBR-2).

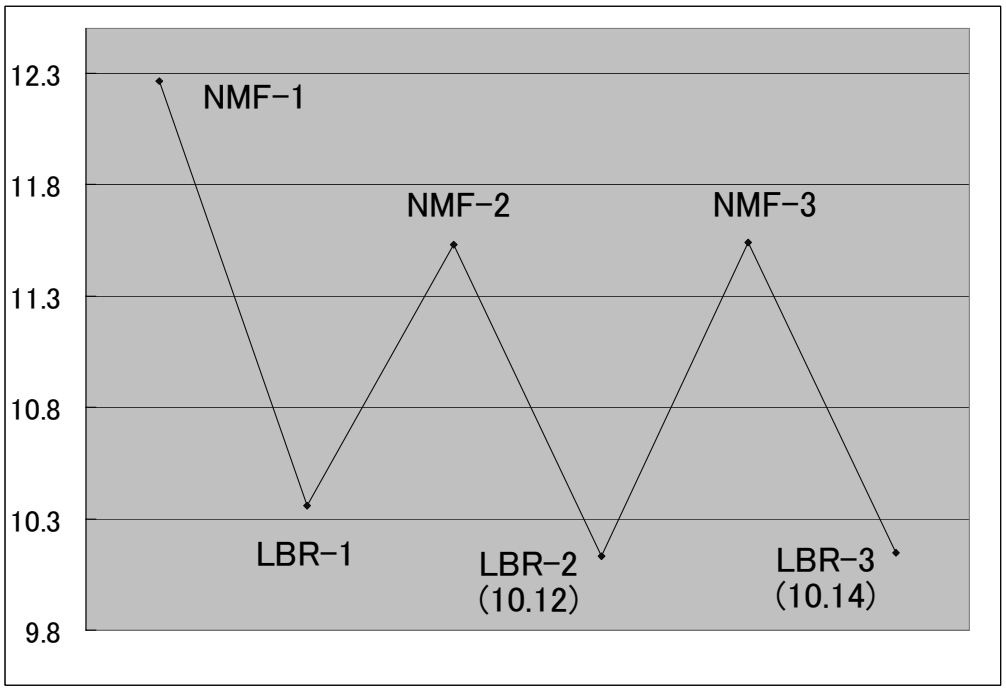

Figure 2: Value of the object function in the ping-pong clustering (2)

In the above example, both of NMF and LBR improve the given clustering result. In this case, the value of the object function of NMF-3 is larger than one of LBR-2. So, we can stop the ping-pong at that time. That is, LBR-3 is needless. However, in many cases, NMF cannot improve the value of the object function and the actual accuracy of clustering. For example, Figure 2 shows the result of our ping-pong clustering for the data set 'kb1' used in our experiment.
In this case, we stop the ping-pong after comparing LBR-2 and LBR-3, and output the clustering result (LBR-2). As shown Figure 2, NMF-2 is poorer than LBR-1. However, NMF-2 is better than NMF-1. Furthermore, LBR-2, which is improved from NMF-2, is better than LBR-1. That is, it is not the good strategy to stop the ping-pong by evaluating the clustering result produced by NMF. Our ping-pong clustering aims to handle the case like the Figure 2 by devising the stop condition of the ping-pong.

5. Experiment

In experiments, we use 16 data sets provided in the following CLUTO site:

```
http://glaros.dtc.umn.edu/gkhome/
      cluto/cluto/download
```

In each data set, the document vector is not normalized. We normalize them by TF-IDF.
For data sets, we conduct four types clustering methods, (1) k-means, (2) NMF, (3) LBR after NMF (NMF+LBR) and (4) our method (Ping-Pong). The difference between NMF+LBR and Ping-Pong is with or without the ping-pong process. NMF+LBR does not pass the clustering result produced by LBR to NMF, that is, it is without the ping-pong process. On the other hand, Ping-Pong does it.
The Table 2 shows the result of the experiment. "KM", "NMF", "NMF+LBR" and "PP(NMF)" mean the result of k-means, NMF, NMF+LBR and our method respectively.
The value in the table means the entropy. The entropy is an evaluation measure for the clustering result. Let $\{K_h\}_{h=1}^k$ and $\{C_j\}_{j=1}^k$ to be the golden answer for the clustering and the clustering result respectively. The entropy E_i of the cluster C_i is defined as follows:

$$E_i = -\sum_{h=1}^{k} P(K_h|C_i) \log P(K_h|C_i)$$

The probability $P(K_h|C_i)$ is estimated by

$$\frac{|K_h \cap C_i|}{|C_i|}.$$

We can get the entropy by taking the weighed mean of a set of $\{E_1, E_2, \cdots, E_k\}$ with weights $\{w_1, w_2, \cdots, w_k\}$ where w_i is ratio of the number of data in C_i to the number N of whole data. That is, the the entropy of $\{C_j\}_{j=1}^k$ is defined by

$$\sum_{i=1}^{k} w_i E_i = -\sum_{i=1}^{k} \frac{|C_i|}{N} \sum_{h=1}^{k} \frac{|K_h \cap C_i|}{|C_i|} \log \frac{|K_h \cap C_i|}{|C_i|}$$

The smaller the entropy is, the better the clustering result is.
The Table 2 shows the effectiveness of our method.

Table 1: Document data sets

Data	# of documents	# of terms	# of classes
cranmed	2431	41681	2
fbis	2463	2000	17
hitech	2301	126373	6
k1a	2340	21839	20
k1b	2340	21839	6
la1	3204	31472	6
la2	3075	31472	6
re0	1504	2886	13
re1	1657	3758	25
reviews	4069	126373	5
tr12	313	5804	8
tr23	204	5832	6
tr31	927	10128	7
tr41	878	7454	10
tr45	690	8261	10
wap	1560	6460	20

Table 2: Experiment result

Data	KM	NMF	NMF+LBR	PP(NMF)
cranmed	0.106	0.748	0.067	**0.055**
fbis	**0.330**	0.383	0.360	0.358
hitech	**0.597**	0.724	0.678	0.679
k1a	0.403	0.384	0.370	**0.352**
k1b	0.306	0.277	0.233	**0.218**
la1	0.660	0.547	0.430	**0.401**
la2	0.620	0.565	**0.411**	0.413
re0	0.384	0.397	**0.373**	0.386
re1	0.391	0.355	**0.310**	0.316
reviews	0.406	0.602	**0.323**	**0.323**
tr12	0.641	0.424	0.406	**0.357**
tr23	0.484	0.473	**0.382**	0.399
tr31	0.373	0.393	0.327	**0.310**
tr41	0.381	0.269	0.277	**0.242**
tr45	0.473	0.254	**0.210**	0.247
wap	0.427	0.378	0.378	**0.371**
Average	0.436	0.448	0.346	**0.339**

6. Discussions

The constitutive method of the ping-pong clustering needs following two conditions.

C1 The input is corresponding to a clustering result.

C2 The output is improved from the input.

Our ping-pong clustering uses NMF and LBR as the constitutive method. Both methods satisfy the condition C1. However the condition C2 is not always satisfied in both methods.

Comparing NMF and NMF+LBR, NMF+LBR has lower entropies than NMF in 14 out of 16 data sets, has the equal entropies in a data set 'wap', and has the higher entropy than NMF in only one data set 'tr41.' Thus, this means that LBR almost satisfies the condition C2.

Next we checked whether NMF improved the clustering result passed by the first LBR. Entropies are reduced for 5 out of 16 data sets, and increased for the rest 11 data sets. This result means hat the NMF does not often satisfy the condition C2.

This problem is caused by the object function (Eq.4) of NMF. The iteration of NMF algorithm improves the value of Eq.4 monotonically. However, the improvement of Eq.4 does not always mean the improvement of the clustering result. This problem is just discussed in the paper (Shinnou and Sasaki, 2007). In this cause, it is hard to use NMF in the ping-pong clustering. To handle this problem, we devise the stop condition of the ping-pong.

We judge whether the ping-pong is stopped or kept, by evaluating only the clustering result produced by LBR. Therefore, even if NMF does not improve the given clustering result, the negative effect for the final clustering result is little.

It is our future work to investigate the relation between the input of NMF and the accuracy of clustering.

By the way, k-means is the typical method that we can use as the constitutive method of the ping-pong clustering.

Table 3: Ping-pong clustering using k-means

Data	KM	KM+LBR	PP(KM)	PP(NMF)
cranmed	0.106	0.070	0.070	**0.055**
fbis	0.330	**0.325**	**0.325**	0.358
hitech	**0.597**	0.619	0.613	0.679
k1a	0.403	0.387	0.376	**0.352**
k1b	0.306	0.246	0.240	**0.218**
la1	0.660	0.440	0.425	**0.401**
la2	0.620	0.421	0.421	**0.413**
re0	0.384	0.385	**0.379**	0.386
re1	0.391	0.351	0.330	**0.316**
reviews	0.406	0.358	0.364	**0.323**
tr12	0.641	0.422	**0.321**	0.357
tr23	0.484	0.457	0.457	**0.399**
tr31	0.373	**0.235**	**0.235**	0.310
tr41	0.381	0.312	0.318	**0.242**
tr45	0.473	**0.195**	0.261	0.247
wap	0.427	0.378	**0.361**	0.371
Average	0.436	0.350	0.343	**0.339**

For reference, we tried the ping-pong clustering using k-means and LBR. Table 3 shows that result. In that table, "KM+LBR" and "PP(KM)" mean the result of LBR for the given clustering result produced by k-means and the result of the ping-pong clustering using k-means and LBR respectively.

Table 3 also shows that LBR almost satisfies the condition C2. The difference between NMF and k-means in the ping-pong clustering is subtle. In the above experiment, NMF was a little better than k-means.

However, NMF produces more informative result than k-means. For example, the matrix produced by NMF includes the degree that each data belongs to a cluster and the degree that each word relates to a cluster. If we improve the clustering result more, these information is useful.

In future, we will investigate the relation between the initial value of NMF and accuracy of output, and use matrices produced by NMF in order to improve the clustering result.

7. Conclusion

In this paper, we proposed the new ping-pong clustering using NMF and LBR as constitutive methods, in order to improve the clustering result produced by NMF. Both NMF and LBR do not always improve the given clustering result. In actual, NMF cannot often do it, but LBR can almost do it. We devise the stop condition of the ping-pong to handle with this problem. In the experiment, we compared our method with the k-means and NMF using 16 document data sets, and evaluated clustering results by entropy. Our experiment showed that our method is effective. In future, we will investigate the relation between the initial value of NMF and accuracy of output, and use matrices produced by NMF in order to improve the clustering result.

Acknowledgements

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas, " Japanese Corpus ", 19011001, 2007.

8. References

Inderjit S. Dhillon, Yuqiang Guan, and J. Kogan. 2002. Iterative Clustering of High Dimentional Text Data Augmented by Local Search. In *The 2002 IEEE International Conference on Data Mining*, pages 131–138.

Chris Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst Simon. 2001. Spectral Min-max Cut for Graph Partitioning and Data Clustering. In *Lawrence Berkeley National Lab. Tech. report 47848.*

Chris Ding, Tao Li, and Wei Peng. 2006. Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence, Chi-square Statistic, and a Hybrid Method. In *AAAI National Conf. on Artificial Intelligence (AAAI-06)*.

Marti A. Hearst and Jan O. Pedersen. 1996. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of SIGIR-96*, pages 76–84.

Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, and Raghu Krishnapuram. 2004. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. In *Proceedings of WWW-04*, pages 658–665.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562.

Michael W. Berry, editor. 2003. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer.

Hiroyuki Shinnou and Minoru Sasaki. 2007. Document clustering by Mcut+NMF (in Japanese). In *13th annaul meeting of Natural Language Processing Association*, pages 558–561.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of SIGIR-03*, pages 267–273.