Careful Seeding Method based on Independent Components Analysis for k-means Clustering

Takashi Onoda
System Engineering Lab., CRIEPI, Tokyo, JAPAN. Email: onoda@criepi.denken.or.jp
Miho Sakai
Tokyo Institute of Technology, Yokohama, JAPAN. Email: sakai@ntt.dis.titech.ac.jp
Seiji Yamada
National Institute of Informatics/SOKENDAI/Tokyo Institute of Technology, Tokyo, JAPAN
Email: seiji@nii.ac.jp

***Abstract*— The k-means clustering method is a widely used clustering technique for the Web because of its simplicity and speed. However, the clustering result depends heavily on the chosen initial clustering centers, which are uniformly chosen at random from the data points. We propose a seeding method that is based on the independent component analysis for the k-means clustering method. We evaluate the performance of our proposed method and compare it with other seeding methods by using benchmark datasets. We also applied our proposed method to a Web corpus, which was provided by ODP, and the CLUTO datasets. The results from the experiments showed that the normalized mutual information of our proposed method is better than the normalized mutual information of the k-means clustering method, the KKZ method, and the k-means++ clustering method.**

***Index Terms*— k-means clustering method, KKZ method, k-means++ clustering method, independent components analysis, seeding**

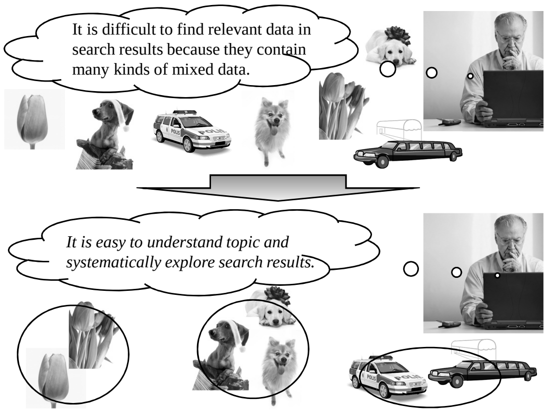
It is difficult to find relevant data in search results because they contain many kinds of mixed data.
It is easy to understand topic and systematically explore search results.

Figure 1. Effect of clustering the results of Web search.

I. Introduction

Clustering is one of the most useful unsupervised learning in data mining[1][2]. It has been applied to various fields and used widespread both in research and business[3]. We are interested in application of clustering to the Web clustering. The Web clustering[4] has a very wide ranges including clustering searched results[5], [6], [7], [8], clustering Web pages/sites[9], [10], [11], clustering Web multimedia[12] and so on. Especially, we focus on clustering of Web searched results because our final research objective is to build IWI (Intelligent Web Interaction) systems. While Web search engines are definitely good for certain search tasks such as finding an organization's Web page, they may be less effective at satisfying ambiguous queries. The results on different subtopics or meanings of the input query also will come together in a hit list, thus implying that the user may have to sift through a large number of irrelevant items to locate those of interest. On the other hand, there is no way to estimate what is relevant to the user given that the queries are usually very short and their interpretation is inherently ambiguous in the absence of context.

An approach for clustering the results of Web search[5], [6], [7], [8] is different from one for retrieving information from the Web. This clustering approach shows the results, which are manually or automatically associated with clusters that consist of similar items (Figure 1).

We consider this clustering of Web searched results should be quick and accurate because a user never wait for the clustered results so long. In particular, we are interested in the simplest and quickest clustering method. Therefore, we deal with the k-means clustering method in our research. We particularly discuss how to solve the problem of "seeding" in the k-means clustering method[13], [14], [15].

The rest of this paper is organized as follows. Section II discusses the related work and the k-means clustering, the KKZ clustering, and the k-means++ clustering methods. Section III discusses the problem with these clustering methods and introduces our proposed method. Section IV presents our experimental results along with comparison of the performance of the proposed method with those of the k-means clustering, the KKZ clustering, and k-means++ clustering methods. Section V concludes this research.

doi:10.4304/jetwi.4.1.51-59

II. Related Works

Clustering is a classic problem in machine learning and computational geometry. In the popular k-means formulation, one is given an integer k and a set of n data points $\mathbf{X} \subset \mathbf{R}^m$. k is the number of cluster centers. The goal is to choose k centers $\mathcal{C}$ to minimize the sum of the squared distances between each point and its closest center.

$$\phi = \sum_{\mathbf{x}\in\mathbf{X}} \min_{\mathbf{c}\in\mathcal{C}} \|\mathbf{x}-\mathbf{c}\|^2 \tag{1}$$

Solving this problem is NP-hard, even with just two clusters[16], however Lloyd[17] proposed a local search solution 25 years ago that is still widely used today.

In this section, we formally define the k-means clustering method, the KKZ clustering method and the k-means++ clustering method.

A. k-means clustering method

The k-means clustering method is simple and fast and locally improves the centers of mass of clusters. It works as follows.

1) Arbitrarily choose k initial centers $\mathcal{C} = \mathbf{c}_1, \cdots, \mathbf{c}_k$,
2) For each $i \in \{1, \ldots, k\}$, set the cluster $\mathbf{c}_i$ to be the set of points in $\mathbf{X}$ that are closer to $\mathbf{c}_i$ than they are to $\mathbf{c}_j$ for all $j \neq i$.
3) For each $i \in \{1, \ldots, k\}$, set $\mathbf{c}_i$ to be the center of the mass of all the points in a set C_i of cluster i: $\mathbf{c}_i = \frac{1}{|C_i|}\sum_{\mathbf{x}\in C_i}\mathbf{x}$.
4) Repeat steps 2) and 3) until $\mathbf{c}_i$ no longer changes.

It is standard practice to uniformly choose the initial centers at random from $\mathbf{X}$. For Step 2), the ties may be arbitrarily broken, as long as the method is consistent. Steps 2) and 3) are both guaranteed to decrease ϕ; therefore, the method makes local improvements to an arbitrary cluster until it is no longer possible to do so.

The k-means method is attractive in practice because it is simple and generally fast. Unfortunately, it is guaranteed only to find a local optimum, which can often be quite poor.

B. KKZ clustering method

The KKZ method was proposed by Katsavounidis et al. [18]. This method calculates the entire distance among the data and finds the data with a wide distance. The data are selected as the initial cluster centers. At any given time, let $D(\mathbf{x})$ denote the shortest distance from a data point $\mathbf{x}$ to the closest center we have already chosen. Then, the following clustering method is defined as the KKZ clustering method[18].

1a) Choose initial centers $\mathbf{c}_1$ and $\mathbf{c}_2$. The distance between $\mathbf{c}_1$ and $\mathbf{c}_2$ is the widest of all distance between a data point and the other data point (Figure 2).
1b) For all data, $D(\mathbf{x}_j), j \in \{1, \cdots, n\}$ are calculated (Figure 3).

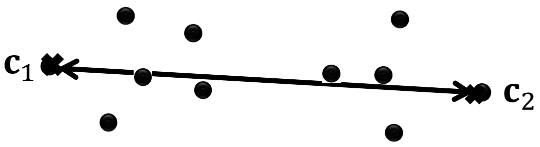

Figure 2. Initial centers of KKZ method

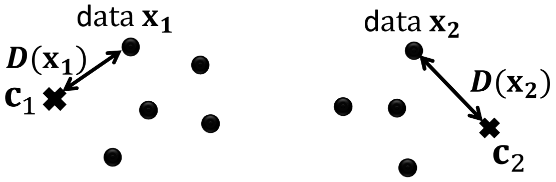

Figure 3. Distance $D(\mathbf{x})$

1c) Choose the next center $\mathbf{c}_i$, selecting $\mathbf{c}_i = \mathbf{x}' \in \mathbf{X}$ with the widest distance $D(\mathbf{x}')$ (Figure 4).
1d) Repeat step 1b) until we have chosen a total of k centers.

Steps 2)-4) proceed just like that for the standard k-means algorithm.

The KKZ method is attractive in practice because it is simple for decision of unique initial centers. However, the KKZ method sometimes find bad clusters because unfortunately it depends on outlier data points.

C. k-means++ clustering method

The k-means method begins with an arbitrary set of cluster centers. k-means++ clustering proposes specifically choosing these centers. At any given time, let $D(\mathbf{x})$ denote the shortest distance from a data point $\mathbf{x}$ to the closest center we have already chosen. Then, the following clustering method is defined as the k-means++ clustering method[19].

1a) Choose an initial center $\mathbf{c}_1$ uniformly at random from $\mathbf{X}$.
1b) For all data, $D(\mathbf{x}_j); j \in \{1, \cdots, n\}$ are calculated (Figure 3).
1c) Randomly generate a real value L satisfying the following equation.

$$0 < L \leq \sum_{\mathbf{x}\in\mathbf{X}} D(\mathbf{x}_j) \tag{2}$$

1d) Choose the next center $\mathbf{c}_i$, selecting the $\mathbf{c}_i = \mathbf{x}_j$ with satisfying the following equation (Figure 5).

$$\sum_{m=1}^{j-1} D(\mathbf{x}_m) < L \leq \sum_{m=1}^{j} D(\mathbf{x}_m) \tag{3}$$

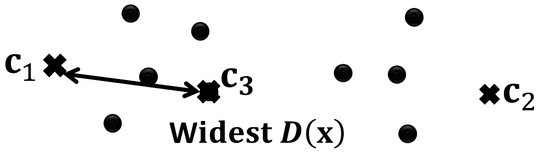

Figure 4. Next center $\mathbf{c}_i$ of KKZ method

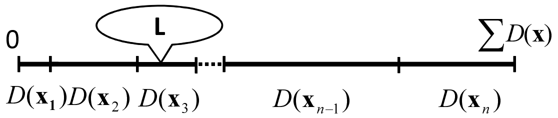

Figure 5. Next center $\mathbf{c}_i$ of k-means++ method

1e) Repeat step 1b) until we have chosen a total of k centers.

Steps 2)-4) proceed in the same way as with the standard k-means clustering method. We call the weighting used in Step 1b) simply "D^2 weighting".

III. Proposed Method

This section describes a problem with the k-means and the k-means++ clustering methods. Then, we propose a k-means combined with an Independent Component Analysis (ICA)[20], [21], [22] based seeding method.

A. Problem for k-means and k-means++ clustering methods

We have six data points, which consist of $\mathbf{x}_i$ (i =1, . . . , 6) and these points are divided into two clusters. Figure 6 shows these six data points.

In addition, Figure 7 shows the global optimal clustering result for these six data points. The first cluster consists of $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}$ and the other consists of $\{\mathbf{x}_3, \mathbf{x}_6\}$. We assume that most of clustering methods can find the global optimal clusters. However, the k-means clustering method generates bad clusters if $\mathbf{x}_2$ and $\mathbf{x}_5$ are chosen as the initial $\mathbf{c}_1$ and $\mathbf{c}_2$ cluster centers. Figure 8 shows the local optimal clusters, which are bad clusters. The k-means++ clustering method was developed to avoid this bad clustering.

However, the k-means++ clustering method sometimes generates bad clusters because it depends on the choice of the initial center $\mathbf{c}_1$. The initial center $\mathbf{c}_1$ is chosen uniformly at random from $\mathbf{X}$.

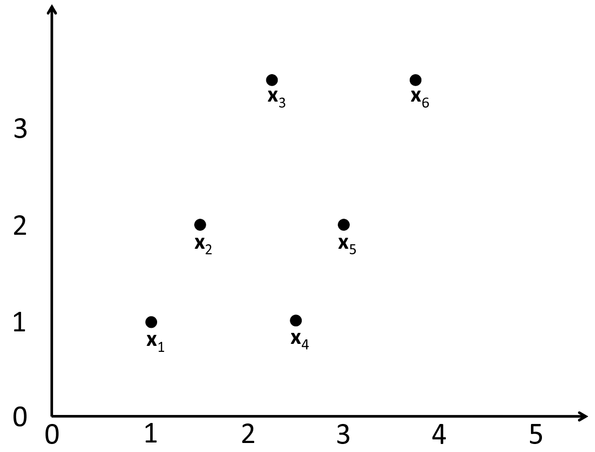

Figure 6. Given Data

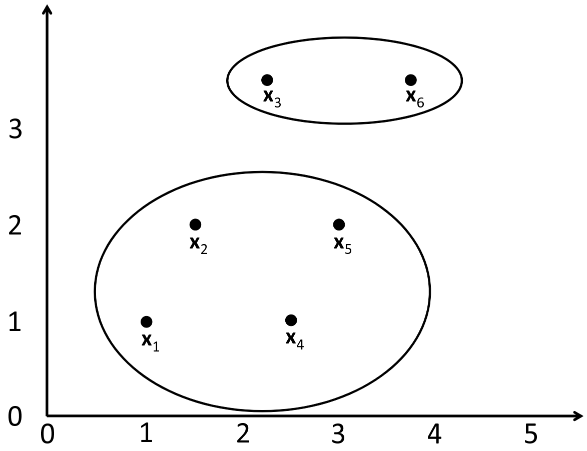

Figure 7. Global Optimal Clustering Case

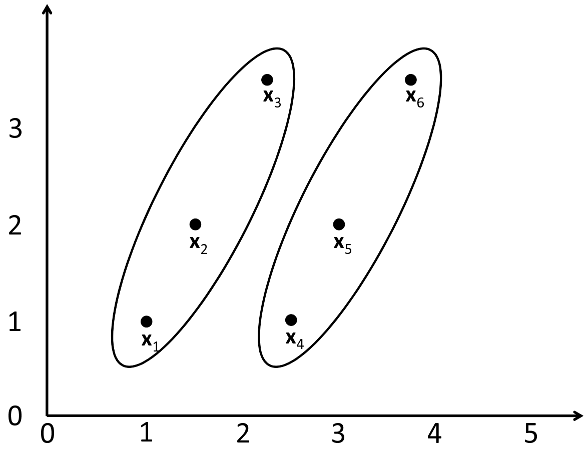

Figure 8. Local Optimal Clustering Case

B. k-means combined with ICA based seeding method

The k-means clustering method begins with an arbitrary set of cluster centers. The k-means++ clustering method begins with a small arbitrary set of cluster centers. As stated above, we propose a method for specifically choosing these centers. At any given time, we can obtain independent components (ICs) from given data $\mathbf{X}$. Then, we define the following seeding method.

1a) Extract k independent components $\mathbf{IC}_1, \cdots, \mathbf{IC}_k$ from given data $\mathbf{X}$ (Figure 9).

1b) Choose k initial centers $\mathbf{c}_i$ ($i = 1, \cdots, k$), selecting $\mathbf{c}_i = \mathbf{x}' \in \mathbf{X}$ with a minimum $\frac{\mathbf{IC}_i \cdot \mathbf{x}'}{|\mathbf{IC}_i||\mathbf{x}'|}$ (Figure 10).

Steps 2)-4) proceed in the same way as with the standard k-means clustering method. Figure 11 shows the concept of the k-means clustering method combined with the ICA based seeding method. In the figure 11, $\mathbf{IC}_1$ and $\mathbf{IC}_2$ denote independent components. The each independent component may become an initial seed to generate the global optimal clustering case.

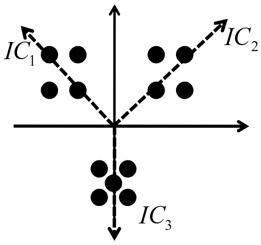

Figure 9. Independent components

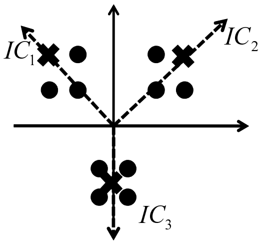

Figure 10. Initial centers $\mathbf{c}_i$ of ICA based method

IV. Experimental condition

To evaluate the k-means clustering method, KZZ method, k-means++ clustering method and proposed method in practice, we implemented and tested them in matlab. In this section, we briefly explain about the data sets that were used for the experiments, the evaluation metrics in the experiments, some compared seeding methods and the results of the experiments. We found that the k-means clustering method combined with the ICA based seeding method performed well in the experiments.

A. Data sets

We evaluated the performance of the k-means clustering method, KKZ method, k-means++ clustering method, and the proposed method using on two kinds of data sets. One contained a small amount of data and the other a large data. The small data set consisted of the UCI Machine Learning repository and Open Directory Project(ODP) Web corpus. The large data set consisted of the CLUTO data sets.

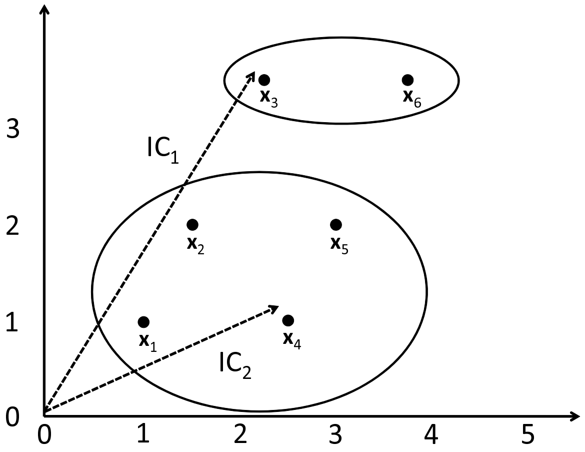

Figure 11. Concept of Our Proposed Method

TABLE I.
No. of clusters, attributes, and samples for UCI repository data sets

Data set	No. of clusters	No. of attributes	No. of samples
Iris	3	4	150
Wine	3	13	178
Soybean -Small	4	35	47
Breast -Cancer	2	9	683

TABLE II.
No. of directories, attributes and samples for ODP corpus data set

Data set	No. of clusters	No. of attributes	No. of samples
ODP	4	340	72

1) UCI Machine Learning repository: The UCI Machine Learning repository had four data sets in our experiments. The first data set, *iris*, consisted of 50 samples from each of three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). The second data set, *wine*, contained the results of a chemical analysis on wines produced in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wine. The third data set, *soybean-small*, was for diagnosing four soybean diseases. The data set consisted of 47 samples and 35 attributes. The fourth data set, *breast-cancer*, contained diagnosis results of breast cancer. The data set consisted of 683 samples and 9 attributes. Table I lists the numbers of samples, the numbers of attributes and the numbers of clusters in the data sets used in our experiments.

2) ODP corpus data: We used the ODP Web corpus data set for our test experiment. The ODP Web corpus was extracted from the Open Directory Project[1] by ourselves. We selected twelve subdirectories from the “Science” top directory, and downloaded top pages of the web sites listed in each directory. We removed tags and stopwords from the pages, and stemmed each word. The summary of each directory is listed in Table I. We treated each directory as a target cluster, and made four datasets using those clusters. Table II lists the number of samples, the number of attributes, and the number of directories of the data sets used in our experiments.

3) CLUTO data sets: CLUTO[2] is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of various clusters. In our experiments, seven CLUTO data sets were adopted. The seven CLUTO data sets are *tr11*, *tr12*, *tr31*, *tr41*, *tr45*, *k1b*, and *re1*. Table III describes statistics on CLUTO

[1] http://www.dmoz.org/
[2] http://glaros.dtc.umn.edu/gkhome/views/cluto

TABLE III.
NO. OF CLUSTERS, ATTRIBUTES AND SAMPLES FOR CLUTO DATA SETS

Data set	No. of clusters	No. of attributes	No. of samples
tr11	9	6429	414
tr12	8	5804	313
tr31	7	10128	927
tr41	10	7454	878
tr45	6	8261	690
k1b	6	21839	2340
re1	25	3758	1657

data sets including the number of samples, the number of attributes and the number of clusters.

B. Evaluation Metrics

We used normalized mutual information as a metric to evaluate the qualities of the clustering outputs from the different methods. The normalized mutual information measures the consistency of the clustering output compared to the ground truth. It reaches a maximum value of 1 only if the membership ϕ_c perfectly matches ϕ_g and a minimum of zero if the assignments of ϕ_c and ϕ_g are independent. The membership function $\phi_c(\mathbf{x})$ is the mapping of a point $\mathbf{x}$ to one of the k clusters. The membership $\phi_g(\mathbf{x})$ represents the true cluster label for $\mathbf{x}$. Formally, the normalized mutual information is derived using the following equation

$$NMI(\phi_g, \phi_c) = \frac{MI(\phi_g, \phi_c)}{\max(H(\phi_g), H(\phi_c))} \tag{4}$$

where $MI(\phi_g, \phi_c)$ denotes the next equation

$$MI(\phi_g, \phi_c) = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{g,c}(i,j) \log \frac{p_{g,c}(i,j)}{p_g(i) p_c(j)} \tag{5}$$

$H(\phi_g)$ comes from the following equation

$$H(\phi_g) = \sum_{i=1}^{k} p_g(i) \log \frac{1}{p_g(i)} \tag{6}$$

and $H(\phi_c)$ denotes the next equation

$$H(\phi_c) = \sum_{j=1}^{k} p_c(j) \log \frac{1}{p_c(j)} \tag{7}$$

The $p_g(i)$ is the percentage of points in cluster i based on the ground truth, i.e.

$$p_g(i) = \frac{\sum_{l=1}^{n} 1(\phi_g(\mathbf{x}_l) - i)}{n}. \tag{8}$$

Similarly, $p_c(j)$ denotes the following equation

$$p_c(j) = \frac{\sum_{l=1}^{n} 1(\phi_c(\mathbf{x}_l) - j)}{n} \tag{9}$$

and $p_{g,c}(i,j)$ is the percentage of points that belong to cluster i in ϕ_g and also cluster j in ϕ_c, i.e.

$$p_{g,c}(i,j) = \frac{\sum_{l=1}^{n} 1(\phi_g(\mathbf{x}_l) - i) 1(\phi_c(\mathbf{x}_l) - j)}{n}. \tag{10}$$

The above defined metrics were used to evaluate the accuracy of the k-means clustering method, KKZ method, k-means++ clustering method and the proposed methods.

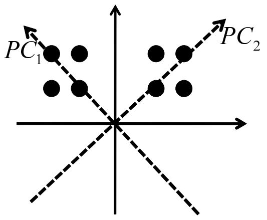

Figure 12. Principal components

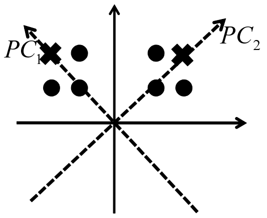

Figure 13. Initial centers $\mathbf{c}_i$ of PCA based method

C. Compared methods

In our experiments, we tried to compare the performance of our proposed methods with the performance of other methods. The other methods consist of the k-means clustering method, KKZ method, and k-means++ clustering method (See section II). Our proposed methods are based on a k-means combined with an ICA based seeding method and k-means combined with a PCA[23] based seeding method. k-means combined with ICA based seeding method was explained in Section III. Now, we briefly explain the k-means combined with a PCA based seeding method.

At any given time, we can obtain principal components (PCs) from the given data $\mathbf{x}$. Then, we define the following seeding method.

1a) Extract k principal components $\mathbf{PC}_1, \ldots, \mathbf{PC}_k$ from given data $\mathbf{X}$ (Figure 12).

1b) Choose k initial centers $\mathbf{c}_i$ ($i = 1, \cdots, k$), selecting $\mathbf{c}_i = \mathbf{x}' \in \mathbf{X}$ with a minimum $\frac{\mathbf{PC}_i \cdot \mathbf{x}'}{|\mathbf{PC}_i||\mathbf{x}'|}$ (Figure 13).

Steps 2)-4) proceed in the same way as with the standard k-means clustering method. Figure 14 shows the concept of the k-means clustering method combined with a PCA based seeding method.

V. EXPERIMENTS

This section discusses some of the experimental results under the above experimental condition.

A. Experimental results for small data sets

The k-means and k-means++ clustering methods were executed 100 times using different initializations over all four data sets from the UCI repository[3]. In our experiments, the Euclid distance was used as a similarity measure when the k-means clustering method was applied to the UCI repository. The KKZ method and the

[3] http://archive.ics.uci.edu/ml/

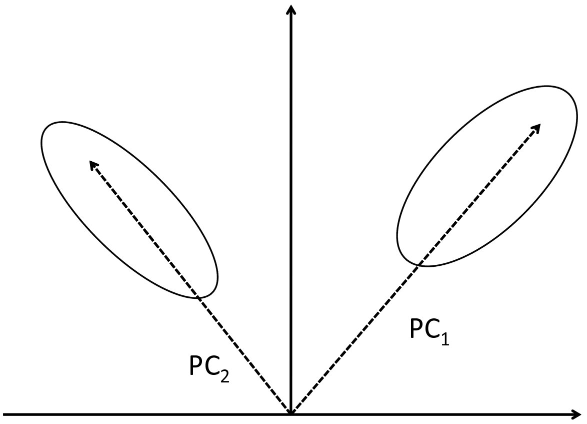

Figure 14. Concept of Our Proposed Method based on PCA

TABLE IV.
EXPERIMENTAL RESULTS FOR *iris* DATA SET

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.751	0.751	0.532	0.703
k-means++	-	0.751	0.751	0.532	0.749
KKZ	0.751	-	-	-	-
PCA	0.751	-	-	-	-
ICA	0.751	-	-	-	-

proposed method were executed only one time because a unique initial seeding can be set up. Table IV lists the experimental results for the *iris* data set. Table V lists the experimental results for the *wine* data set. Table VI lists the experimental results for the *soybean-small* data set. Table VII lists the experimental results for the *breast-cancer* data set. Tables IV, V, VI, and VII have an averaged *NMI*, a maximum *NMI*, a minimum *NMI*, and a *NMI* when the clusters achieved minimum variance.

In IV, V, and VII tables, the *NMI*s of our proposed method are the same as the maximum *NMI*s of the k-means clustering method and the k-means++ clustering method. The *NMI*s of our proposed methods are achieved by using only one initial seeding. Therefore, IV, V, and VII tables show that the proposed method outperforms both the k-means clustering method and the k-means++ clustering method for the *iris*, *wine*, and *breast-cancer* data sets of the UCI repository.

TABLE V.
EXPERIMENTAL RESULTS FOR *wine* DATA SET

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.429	0.429	0.387	0.418
k-means++	-	0.429	0.429	0.387	0.418
KKZ	0.387	-	-	-	-
PCA	0.429	-	-	-	-
ICA	0.429	-	-	-	-

TABLE VI.
EXPERIMENTAL RESULTS FOR *soybean-small* DATA SET

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.711	1.000	0.518	0.714
k-means++	-	0.711	1.000	0.711	0.806
KKZ	0.711	-	-	-	-
PCA	0.711	-	-	-	-
ICA	0.711	-	-	-	-

TABLE VII.
EXPERIMENTAL RESULTS FOR *breast-cancer* DATA SET

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.743	0.743	0.743	0.743
k-means++	-	0.743	0.743	0.743	0.743
KKZ	0.743	-	-	-	-
PCA	0.743	-	-	-	-
ICA	0.743	-	-	-	-

TABLE VIII.
EXPERIMENTAL RESULTS FOR *ODP Web corpus* DATA SET

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.555	0.589	0.392	0.514
k-means++	-	0.555	0.589	0.425	0.525
KKZ	0.531	-	-	-	-
PCA	0.500	-	-	-	-
ICA	0.638	-	-	-	-

In IV and VII tables, the *NMI*s of the KKZ method are the same as the maximum *NMI*s of the k-means clustering method and the k-means++ clustering method. The *NMI*s of the KKZ method are achieved by using only one initial seeding. Therefore, the IV and VII tables also show that the KKZ method outperforms both the k-means clustering method and the k-means++ clustering method for the *iris* and *breast-cancer* data sets of the UCI repository. However the V table shows that the performance of the KKZ method is the worst among the compared methods in our experiments.

We generally cannot provide true cluster data. Having a *NMI* with minimum variance is the most important issue for real-world applications. Table VI shows that the *NMI*s of our proposed method are the same as the *NMI*s of the k-means clustering method and k-means++ clustering method when the clusters achieved minimum variance. This situation shows that the performance of our proposed method is the same as the performance of the k-means clustering method and the k-means++ clustering method for the *soybean-small* data set. And the *NMI* with minimum variance is achieved by using only one initial seeding.

In our experiments, the k-means clustering and k-means++ clustering methods run 100 times using different initializations for the *ODP Web corpus* data set. The proposed method runs only one time because it can set up a unique initial seeding. Table II lists the experimental results of the *ODP Web corpus* data set. When the k-means clustering method was applied to an ODP corpus, the cosine distance was used as a similarity measure in our experiments. The KKZ method and the proposed method were executed only one time because they can set up a unique initial seeding.

Table VIII shows that the *NMI* of our proposed method is better than the *NMI* of the k-means clustering and k-means++ clustering methods when the clusters achieved a minimum variance for the *ODP Web corpus* data set. Table VIII shows that the NMI of the proposed method was 0.638. This value is better than the maximum NMI of the k-means clustering method and the maximum NMI of k-means++ clustering method. In addition, the NMI of the proposed method is better than the NMI of the KKZ method and the NMI of PCA based method. Therefore, Table II shows that the proposed method outperforms the k-means clustering method, k-means++ clustering method, KKZ method and PCA based method for the *ODP Web corpus* data set. In addition, the best NMI is achieved by using only one initial seeding.

B. Experimental results for large data sets

The k-means and k-means++ clustering methods were executed 100 times using different initializations for all seven data sets from the CLUTO data sets. In our experiments, the cosine distance was used as a similarity measure when k-means clustering method was applied to CLUTO data sets. KKZ method and the proposed method were executed only one time because they could set up a unique initial seeding. Table IX lists the experimental results for the seven CLUTO data sets.

It is difficult to understand aspects of performance of compared methods from table IX. Therefore, we would now like to introduce the ratio between the number of attributes and the number of samples. We could order the results from the CLUTO data sets by using the value of the ratio. Table X lists the ordered results of CLUTO data sets. The underlined performances indicate the best ones. The ratio denotes the following equation

$$ratio = \frac{\text{No. of attributes}}{\text{No. of samples}}. \tag{11}$$

The NMIs of the k-means and k-means++ are NMIs with the minimum variance in table IX.

When the ratio is smaller than 10.00, table X shows that shows the performance the proposed method based on ICA is better than the performance of the k-means clustering method, k-means++ clustering method and KKZ method. In other words, the NMIs of the proposed method is better than the NMIs of the k-means clustering method, k-means++ clustering method and KKZ method for *k1b*, *tr41* and *re1* CLUTO data sets. However, when the ratio is larger than 10.00, the performance of the proposed method based on ICA is not better than the performances of the k-means clustering method, k-means++ clustering method and KKZ method in table X. In other words, the NMIs of the proposed method are not better than the NMIs of k-means clustering method, k-means++ clustering method, and KKZ method for *tr12*, *tr11*, *tr45*, and *tr31* CLUTO data sets. When the ratio is larger than 10.00, the number of attributes is much larger than the number of samples and it is difficult to find stable independent components. Therefore, the proposed method with ICA may not be able to perform well.

TABLE IX.
EXPERIMENTAL RESULTS FOR *CLUTO* DATA SETS

tr11

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.630	0.730	0.523	0.635
k-means++	-	0.669	0.717	0.545	0.632
KKZ	0.578	-	-	-	-
PCA	0.619	-	-	-	-
ICA	0.585	-	-	-	-

tr12

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avrg. *NMI*
k-means	-	0.664	0.752	0.521	0.664
k-means++	-	0.621	0.689	0.425	0.621
KKZ	0.683	-	-	-	-
PCA	0.500	-	-	-	-
ICA	0.638	-	-	-	-

tr31

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.579	0.676	0.377	0.503
k-means++	-	0.523	0.641	0.392	0.507
KKZ	0.439	-	-	-	-
PCA	0.504	-	-	-	-
ICA	0.438	-	-	-	-

tr41

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.623	0.699	0.533	0.611
k-means++	-	0.651	0.730	0.530	0.620
KKZ	0.584	-	-	-	-
PCA	0.680	-	-	-	-
ICA	0.667	-	-	-	-

tr45

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.769	0.787	0.594	0.696
k-means++	-	0.794	0.794	0.564	0.697
KKZ	0.660	-	-	-	-
PCA	0.744	-	-	-	-
ICA	0.722	-	-	-	-

k1b

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.537	0.649	0.421	0.518
k-means++	-	0.523	0.611	0.422	0.521
KKZ	0.491	-	-	-	-
PCA	0.438	-	-	-	-
ICA	0.587	-	-	-	-

re1

method	*NMI*	*NMI* with min. variance	max. *NMI*	min. *NMI*	avg. *NMI*
k-means	-	0.541	0.578	0.460	0.523
k-means++	-	0.545	0.575	0.465	0.545
KKZ	0.557	-	-	-	-
PCA	0.557	-	-	-	-
ICA	0.575	-	-	-	-

C. Computational costs

Next, we explain the computational cost of the proposed method from an experimental point of view. In our experiments, we used a Windows Vista 32 bit machine that has an Intel Core 2 Duo E8600 3.34 GHz and 4 GB memory. Table XI lists computational times of four UCI repository data sets.

We can find from this table that the computational time of the proposed method is larger than the computational time of the k-means clustering method for the four UCI repository data sets. The k-means clustering method was

TABLE X.
ORDERED EXPERIMENTAL RESULTS BASED ON RATIO (NO. OF ATTRIBUTES/NO. OF SAMPLES) FOR *CLUTO* DATA SETS

data set	ratio	proposed	k-means	k-means++	KKZ
tr12	18.54	0.638	0.664	0.621	0.683
tr11	15.52	0.585	0.630	0.669	0.578
tr45	11.97	0.722	0.769	0.794	0.660
tr31	10.92	0.438	0.579	0.523	0.439
k1b	9.33	0.587	0.537	0.523	0.491
tr41	8.49	0.667	0.623	0.651	0.584
re1	2.26	0.575	0.541	0.545	0.557

TABLE XI.
COMPUTATIONAL TIMES (SEC.) FOR UCI REPOSITORY DATA SETS

method	*iris*	*wine*	*breast-cancer*	*soybean-small*
k-means	0.0032	0.0034	0.0059	0.0041
proposed	0.0748	0.0948	0.0643	0.0773

executed 100 times using different initializations for all four data sets of the UCI repository.

Table XII lists computational times of the *tr45 CLUTO* data set. We can find from this table that the computational time of the proposed method is smaller than the computational time of the k-means clustering method for a *CLUTO* data set. The k-means clustering method was executed 100 times using different initializations for all four data sets of the UCI repository. In other words, the computational time of the proposed method is smaller than the computational time of the k-means-clustering method for large data sets that contain many attributes. Generally, the Web contains a lot of documents with many attributes. Therefore, the proposed method is useful for the Web.

TABLE XII.
COMPUTATIONAL TIMES (SEC.) FOR *tr45 CLUTO* DATA SET

method	*tr45*
k-means	90.42×100
proposed	52.39+90.42

VI. CONCLUSION

We proposed a method that combines the k-means clustering method with an Independent Component Analysis based seeding method and a Principal Component Analysis based seeding method, and compared the performances of the proposed method with the performance of the standard k-means clustering method, k-mean++ clustering method, and k-means clustering method with a KKZ seeding method.

From our experimental results for small data sets (UCI repository data sets), our proposed method performed the same as or better than the standard k-means clustering method, k-means++ clustering method, and k-means clustering method with a KKZ seeding method.

From our experimental results for large data sets(CLUTO data sets), our proposed method based on ICA performed better than the standard k-means clustering method, k-means++ clustering method, and k-means clustering method with a KKZ seeding method when the ratio between the number of attributes and the number of samples is smaller than 10.00. When the ratio between the number of attributes and the number of samples is larger than 10.00, our proposed method based on ICA did not perform better than the standard k-means clustering method, k-means++ clustering method, or k-means clustering method with a KKZ seeding method. Generally, the Web has a lot of documents and the ratio between the number of attributes and the number of samples is small. Therefore, the proposed method is useful for the Web.

For our future work, we plan to theoretically analyze the computational cost of the proposed method, and to conduct research on how to decide the number of clusters based on the observed data distribution.

ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 23300063, 2011.

REFERENCES

[1] S. Basu, I. Davidson, and K. Wagstaff, Eds., *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman and Hall/CRC, 2008.

[2] C. Ding, “A Tutorial on Spectral Clustering Part I: Basic Theory,” 2004. [Online]. Available: http://ranger.uta.edu/ chqding/Spectral/

[3] P. Berkhin, “Survey of clustering data mining techniques,” Accrue Software, San Jose, CA, Tech. Rep., 2002.

[4] C. Carpineto, S. Osiński, G. Romano, and D. Weiss, “A survey of web clustering engines,” *ACM Computing Survey*, vol. 41, pp. 17:1–17:38, 2009.

[5] P. Ferragina and A. Gulli, “A personalized search engine based on web-snippet hierarchical clustering,” in *Special interest tracks and posters of the 14th international conference on World Wide Web (WWW’05)*, 2005, pp. 801–810.

[6] R. Navigli and G. Crisafulli, “Inducing word senses to improve web search result clustering,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP’10)*, 2010, pp. 116–126.

[7] B. Stein, T. Gollub, and D. Hoppe, “Beyond precision@10: clustering the long tail of web search results,” in *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM’11)*, 2011, pp. 2141–2144.

[8] Y. Wang and M. Kitsuregawa, “Evaluating contents-link coupled web page clustering for web search results,” in *Proceedings of the eleventh international conference on Information and knowledge management (CIKM ’02)*, 2002, pp. 499–506.

[9] D. Crabtree, P. Andreae, and X. Gao, “Query directed web page clustering,” in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI’06)*, 2006, pp. 202–210.

[10] P. Li, B. Wang, W. Jin, and Y. Cui, “User-related tag expansion for web document clustering,” in *Proceedings of the 33rd European conference on Advances in information retrieval (ECIR’11)*, 2011, pp. 19–31.

[11] C. Lu, X. Chen, and E. K. Park, “Exploit the tripartite network of social tagging for web clustering,” in *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM ’09)*, 2009, pp. 1545–1548.

[12] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma, "Web image clustering by consistent utilization of visual features and surrounding texts," in *Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA'05)*, 2005, pp. 112–121.

[13] T. Onoda, M. Sakai, and S. Yamada, "Independent component analysis based seeding method for k-means clustering," in *In Proceedings of the International Workshop on Intelligent Web Interaction 2011 (IWI-2011)*, 2011, pp. 122–125.

[14] T. Onoda, M. Sakai, and S. Yamad, "Careful seeding based on independent component analysis for k-means clustering," in *In Proceedings of the International Workshop on Intelligent Web Interaction 2010 (IWI-2010)*, 2010, pp. 112–115.

[15] T. Onoda, M. Sakai, and S. Yamad, "Seeding method based on independent component analysis for k-means clustering," in *In Proceedings of Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems (SCIS&ISIS-2010)*, 2010, pp. 1306–1309.

[16] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering large graphs via the singular value decomposition," *Maching Learning*, vol. 56, no. 1-3, pp. 9–33, 2004.

[17] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[18] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized lloyd iteration," *IEEE Signal Processing Letters*, vol. 1, no. 10, pp. 144–146, 1994.

[19] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[20] A. Hyv̈rinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[21] A. Hyv̈rinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[22] A. Hyv̈rinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492,, 1997.

[23] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.

Takashi Onoda graduated from International Christian University, Tokyo, Japan in 1986. He received the M.S. degree in nuclear engineering from Tokyo Institute of Technology, Tokyo, Japan in 1988.

He works at Central Research Institute of Electric Power Industry from 1988. He received the Dr. Eng. degree in mathematical engineering from University of Tokyo, Tokyo, Japan in 2000. He worked as a visiting researcher in GMD FIRST in Berlin from September in 1997 to September in 1998. He is a sector leader at Central Research institute of Electric Power Industry and a visiting professor at Tokyo Institute of Technology.

His research interests are in statistical learning theory and its applications. He is a member of JSAI.

Miho Sakai graduated from Musashi Institute of Technology, Tokyo, Japan in 2009. She received her MS degree in engineering from Tokyo Institute of Technology in 2011. Her research interest is in clustering method.

Seiji Yamada is a professor at the National Institute of Informatics. Previously he worked at Tokyo Institute of Technology. He received B.S. (1984), M.S. (1986) and the Ph.D. (1989) degrees in artificial intelligence from Osaka University.

His research interests are in the design of intelligent interaction including Human-Agent Interaction, intelligent Web interaction and interactive machine learning. He is a member of IEEE, AAAI, ACM, JSAI, IPSJ and HIS.