J Intell Inf Syst (2011) 37:187–216
DOI 10.1007/s10844-010-0132-5

A graph model for mutual information based clustering

Tetsuya Yoshida

Received: 22 December 2009 / Revised: 6 June 2010 / Accepted: 26 August 2010 /
Published online: 12 September 2010

Abstract We propose a graph model for mutual information based clustering problem. This problem was originally formulated as a constrained optimization problem with respect to the conditional probability distribution of clusters. Based on the stationary distribution induced from the problem setting, we propose a function which measures the relevance among data objects under the problem setting. This function is utilized to capture the relation among data objects, and the entire objects are represented as an edge-weighted graph where pairs of objects are connected with edges with their relevance. We show that, in hard assignment, the clustering problem can be approximated as a combinatorial problem over the proposed graph model when data is uniformly distributed. By representing the data objects as a graph based on our graph model, various graph based algorithms can be utilized to solve the clustering problem over the graph. The proposed approach is evaluated on the text clustering problem over 20 Newsgroup and TREC datasets. The results are encouraging and indicate the effectiveness of our approach.

Keywords Clustering · Mutual information · Graph · Cut

1 Introduction

Clustering is a process of finding a partition of data objects into mutually exclusive and exhaustive groups. The groups are called clusters. The objective is to find clusters of data objects such that data within the same group are similar to each other, while data among different groups are dissimilar. Clustering is a fundamental data processing in various fields, and has been investigated in many research communities,

T. Yoshida (✉)
Graduate School of Information Science and Technology,
Hokkaido University, N-14 W-9, Sapporo 060-0814, Japan
e-mail: yoshida@meme.hokudai.ac.jp

e.g., machine learning, data mining, statistical pattern recognition (Jain et al. 1999; Dempster et al. 1977; Hartigan and Wong 1979).

In this paper we consider data clustering based on mutual information under the framework in Tishby et al. (1999). In this framework, the clustering problem is formalized as a constrained optimization problem with respect to the conditional probability distribution of clusters based on mutual information. Since finding out the globally optimal solution is very difficult due to the non-linearity of mutual information and non-convexity of the problem, several algorithms were proposed to find out approximated solutions (Tishby et al. 1999; Slonim and Tishby 2000; Slonim et al. 2002). Although the original formulation allows the probabilistic or soft assignment of data into clusters as in Pereira et al. (1993),[1] *hard assignment*, i.e., each data can be assigned only to one cluster, is widely utilized in many applications of clustering technique. Thus, we focus on the situation where hard assignment is conducted under the formulation in this paper.

Based on the stationary distribution induced from the problem setting, we propose a function which measures the relevance among data objects under the problem setting. The stationary distribution can be considered as representing the relation between data objects and clusters. We extend this relation to the ones among data objects, and define the function to capture the pairwise relation among data objects as their relevance. By mapping each data object to a vertex and connecting the vertices with edges with their relevance, the entire objects can be represented as an edge-weighted graph in our model. The edge-weighted graph for the specified data objects is called a data graph in our approach.

We show that, in hard assignment, the clustering problem based on mutual information can be approximated as a combinatorial problem over the proposed edge-weighted graph for the data objects when data is uniformly distributed. Representing the entire data objects as a data graph based on our graph model and formalizing the clustering problem over the proposed data graph enable to utilize various graph algorithms to solve the clustering problem over the graph. We demonstrate the effectiveness of the proposed approach by utilizing spectral clustering over the proposed graph model and evaluating it on 20 Newsgroup and TREC datasets. The results are encouraging, especially for the correctness of cluster assignment with respect to the true cluster labels.

This paper is organized as follows. Section 2 briefly explains related work to clarify the context of our approach. Section 3 explains the problem setting of mutual information based clustering. Section 4 explains the details of our graph model and the correspondence between the original clustering problem and a combinatorial problem over the proposed graph model. Section 5 reports the evaluation of the proposed approach. Section 6 gives concluding remarks.

2 Related work

In general, clustering methods can be divided into the following approaches: hierarchical methods (Guha et al. 1998), partitioning methods (Hartigan and Wong

[1] Probabilistic assignment of data object into several clusters is called *soft assignment*.

1979; Ng and Han 2002), density-based methods (Ester et al. 1996), and grid-based methods (Agrawal et al. 1998). Hierarchical methods construct a cluster hierarchy, or a tree of clusters (called a dendrogram), whose leaves are the data points and whose internal nodes represent nested clusters of various sizes (Guha et al. 1998). Partitioning methods return a single partition of the entire data under fixed parameters (number of clusters, thresholds, etc.). Each cluster can be represented by its centroid (Hartigan and Wong 1979) or by one of its objects located near its center (Ng and Han 2002). Density-based methods try to find arbitrary-shaped clusters, which are dense regions of objects compared with other regions (Ester et al. 1996). Grid-based algorithms partition the data space into a finite number of cells (or, grids) to form a grid structure, and conduct clustering over the grids (Agrawal et al. 1998). An overview of various clustering methods is described in Jain et al. (1999).

Since our approach is based on mutual information and graph structure, in the following subsections, we briefly overview related work in information-theoretic approach and graph-theoretic approach for data clustering.

2.1 Information-theoretic approach

Information-theoretic criteria have been widely utilized in statistical analysis, machine learning, and data mining. For instance, AIC (Akaike Information Criterion; Akaike 1973) and MDL (Minimum Description Length principle; Rissanen 1978) have been widely utilized in statistical analysis. For the classification problem, various criteria based on mutual information (Cover and Thomas 2006) have been widely utilized for the construction of decision trees. For instance, ID3 (Quinlan 1986) utilized information gain and C4.5 (Quinlan 1993) utilized information gain ratio.

A data clustering framework based on mutual information was proposed in Tishby et al. (1999). Although the framework in Tishby et al. (1999) is about the clustering of data, not about the representation of data, simultaneous clustering (co-clustering) of both data and representation was proposed in Dhillon et al. (2003) based on mutual information. These approaches deal with the clustering of count data. For categorical data clustering, the relationship between entropy-based criterion and other criteria (including MDL, Kullback–Leibler (KL) divergence, etc.) is shown in Li et al. (2004). In addition, an algorithm based on the minimum entropy criterion was proposed and evaluated.

Graphical models have been utilized to describe problems in pattern recognition and machine learning in terms of discrete random variables (Frey 1998). Based on a combinatorial random variable over a combinatorial set (e.g., a power set or all partitioning of a set), combinatorial Markov random fields (called Comraf) over combinatorial random variables were proposed in Bekkerman et al. (2006). Mutual information is also utilized to define potential functions in Comraf, and it was shown that the frameworks in Tishby et al. (1999) and Dhillon et al. (2003) can be described using Comraf graphs.

2.2 Graph-theoretic approach

When a pairwise relation among data objects is specified, the entire objects can be represented as a graph based on the relation. Furthermore, if similarities or

dissimilarities among objects are also specified, the graph structure can be extended into an edge-weighted graph. In order to define a graph structure based on a pairwise relation among objects, neighborhood graphs (or, proximity graphs) have been widely utilized (Toussaint 2005; Muhlenbach and Lallich 2009; Hacid and Yoshida 2010). These are geometrical structures defined based on the proximity of pairwise objects. An extensive survey of various utilization of proximity graphs is described in Toussaint (2005). For instance, proximity graphs were utilized for clustering (Muhlenbach and Lallich 2009) and indexing of high-dimensional data (Hacid and Yoshida 2010).

Based on a graph representation of the entire objects, various graph-theoretic clustering approaches have been proposed. In general, these can be divided into two approaches: vertex based approaches and edge based approaches. In many algorithms, vertex coloring is conducted in the former, and the objects with the same color are collected into the same cluster. On the other hand, in the latter, edge removal is conducted to disconnect the graph. Connected components after the removal of edges are considered as clusters.

As for the vertex coloring based approach for undirected graphs, a hierarchical agglomerative clustering method was proposed in Guénoche et al. (1991). It conducts 2-coloring of the vertices for a maximum spanning tree of a graph. In Hansen and Delattre (1978), partitioning of a given data items into the specified number of clusters is conceived in terms of the minimal coloring of a graph. Recently, a notion of b-coloring of undirected graphs was proposed in Irving and Manlov (1999). A graph b-coloring is a vertex coloring, and it satisfies the following constraints: (i) adjacent vertices have different colors, (ii) in each color, at least one vertex is adjacent to all the other colors. A clustering framework based on the notion of b-coloring is proposed in Elghazel et al. (2008). This approach is also extended to seek for better partitioning (Elghazel et al. 2007; Ogino and Yoshida 2010).

As for edge based approach, Zahn proposed a minimum spanning tree (MST) based divisive algorithm (Zahn 1971). In this approach, edges with large distance in MST are removed to construct disconnected components. Other types of proximity graphs (e.g., Gabriel graphs and relative neighborhood graphs) are dealt with in Urquhart (1982). Removal of edges based on the notion of minimum cut in graph theory (Diestel 2006) is also proposed in Hartuv and Shamir (2000). In this approach, highly connected subgraphs in an undirected graph can be identified with provably good properties. Recently, approximation of minimum cut for identifying clusters have been intensively studied. Various algorithms have been proposed based on spectral graph theory (Chung 1997) for finding out approximated solutions based on the eigenvalues and eigenvectors of a graph (Belkin and Niyogi 2002; von Luxburg 2007).

3 Problem settings

3.1 Preliminaries

Basically, we use a bold italic capital letter to denote a set. Let $\boldsymbol{X}$ be a set of data objects. For a set $\boldsymbol{X}$, $|\boldsymbol{X}|$ represents its cardinality.

Suppose X stands for a random variable over the domain $\mathcal{X}$, and $p_1(x)$ and $p_2(x)$ are probability distributions for X.

Definition 1 Kullback–Leibler (KL) divergence between two probability distributions $p_1(x)$ and $p_2(x)$ for a random variable X is defined as (Cover and Thomas 2006):

$$D_{KL}[p_1(x)||p_2(x)] = \sum_x p_1(x) \log \frac{p_1(x)}{p_2(x)} \tag{1}$$

Suppose X and Y are two random variables (their domains are $\mathcal{X}$ and $\mathcal{Y}$), and $p(x, y)$ stands for their joint probability distribution. Let $p(x)$ and $p(y)$ stand for their marginal probability distributions, and $p(y|x)$ stands for the conditional probability distribution of Y given the observation of X.

Definition 2 Mutual Information $I(X; Y)$ between two random variables X and Y is defined as:

$$I(X; Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{2}$$

$$= D_{KL}[p(x, y)||p(x)p(y)] \tag{3}$$

3.2 The information bottleneck framework

Data clustering based on mutual information was proposed in Tishby et al. (1999). By introducing a relevant random variable Y, the objective is to find clusters $\boldsymbol{T}$ of data objects $\boldsymbol{X}$ such that the clusters are still informative about Y. Random variables X and T corresponds to $\boldsymbol{X}$ and $\boldsymbol{T}$, and T should be completely defined given X and irrelevant to Y. These relation among random variables is formalized as follows.

Definition 3 (IB Markovian relation) The following Markovian relation holds among the random variables X, Y and T:

$$T \leftrightarrow X \leftrightarrow Y \tag{4}$$

where X is a random variable, Y is the relevant variable, and T is a compressed representation of X.

For instance, suppose a set of documents $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ is specified, each of which contains a "bag" of terms to describe the document. Here, the set of whole terms utilized to describe the documents corresponds to $\boldsymbol{Y} = \{y_1, \ldots, y_m\}$. Random variables X and Y correspond to $\boldsymbol{X}$ and $\boldsymbol{Y}$. $p(x, y)$ represents the joint probability of a document x containing a term y, and can be estimated by the co-occurrence of x and y. The goal of data clustering is to find a partition $\boldsymbol{T} = \{t_1, \ldots, t_k\}$ of $\boldsymbol{X}$ such that T is still informative about Y. Here, each $t \in \boldsymbol{T}$ corresponds to a cluster of documents.

Data clustering is formalized as a constrained optimization problem for the conditional probability $p(t|x)$, where t corresponds to a cluster and x corresponds to an object (Tishby et al. 1999).

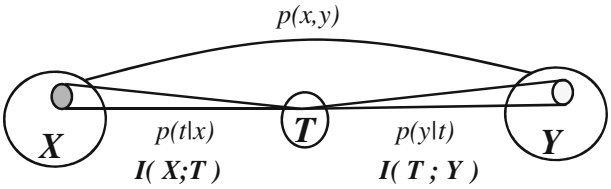

Fig. 1 Data clustering based on mutual information (Tishby et al. 1999)

Problem 1 Find the conditional probability distribution $p(t|x)$ which minimizes the following objective function $\mathcal{L}$

$$\mathcal{L} = I(X; T) - \beta I(T; Y) \tag{5}$$

where $I(X; T)$ is mutual information between X and T, and $I(T; Y)$ is the mutual information between T and Y. β is a control parameter, and corresponds to the inverse temperature in statistical physics.

Intuitively, $I(X; T)$ (the first term in (5)) corresponds to the compactness of new representation T for representing data objects. On the other hand, $I(T; Y)$ (the second term in (5)) represents to what extent the random variable T has information about the random variable Y (and vice versa). Thus, $I(T; Y)$ corresponds to the accuracy of T for predicting the value of relevant variable. Since the objective is the minimization of $\mathcal{L}$ in (5), this is realized by minimizing $I(X; T)$ and by maximizing $I(T; Y)$. Minimizing $I(X; T)$ corresponds to seeking for a compact representation T for data objects, while maximizing $I(T; Y)$[2] corresponds to seeking for a representation T which is still informative about the relevant variable. Problem 1 is illustrated in Fig. 1 (Tishby et al. 1999).

It was shown that the optimal solution of Problem 1 should satisfy the following self-consistent equations (Tishby et al. 1999; Slonim 2002).

Theorem 1 *When $p(x, y)$ and β are specified, and Markovian relation (4) holds, $p(t|x)$ is a stationary point of $\mathcal{L}$ if and only if $p(t|x)$ satisfies the following equations:*

$$p(t|x) = \frac{p(t)}{Z(x, \beta)} \exp(-\beta D_{KL}[p(y|x)||p(y|t)]) \tag{6}$$

$$Z(x, \beta) = \sum_t p(t) \exp(-\beta D_{KL}[p(y|x)||p(y|t)]) \tag{7}$$

where $Z(x, \beta)$ is a normalization term.

3.3 Approximation algorithms

Equation (6) in Theorem 1 indicates that $p(t|x)$ is the stationary distribution under the problem setting. However, $p(t|x)$, the left hand side of (6), implicitly and non-linearly affects its right hand side. Furthermore, the objective function $\mathcal{L}$ in (5) is not

[2]Minimizing $-I(T; Y)$ is equivalent to maximizing $I(T; Y)$.

convex with respect to $p(t|x)$, $p(t)$, $p(y|t)$ simultaneously. Thus, it is quite difficult to find the global optimal solution of Problem 1.

Several algorithms were proposed to find out approximated solutions of (5) (Tishby et al. 1999; Slonim and Tishby 2000; Slonim et al. 2002; Slonim 2002). For instance, the following algorithms were proposed:

- iIB: an *iterative* projection based algorithm (Tishby et al. 1999)
- dIB: a *deterministic* annealing-like algorithm (Slonim 2002)
- aIB: an *agglomerative* algorithm (Slonim and Tishby 2000)
- sIB: a *sequential* re-assignment algorithm (Slonim et al. 2002)

The stationary distribution in (6) is utilized in the iIB algorithm. This is realized by iterating the following equations:

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} \exp(-\beta D_{KL}[p(y|x)||p(y|t)]) \tag{8}$$

$$p(t) = \sum_x p(x)p(t|x) \tag{9}$$

$$p(y|t) = \frac{1}{p(t)} \sum_x p(x,y)p(t|x) \tag{10}$$

where (8) corresponds to the stationary distribution in (6). Iterating over these three equations corresponds to the projection of probability distribution as in EM algorithm (Dempster et al. 1977). Note that different initializations can lead to different solutions in iIB algorithm. Based on the iIB algorithm, dIB algorithm conducts *deterministic* annealing of the parameter β by gradually modifying its value.

In aIB and sIB algorithms, the following objective function is considered, which is a dual form of the original objective function:

$$\mathcal{L}_{\max} = I(T;Y) - \beta^{-1} I(X;T) \tag{11}$$

Based on the objective function in (11), aIB algorithm conducts hierarchical agglomerative clustering by merging two clusters with the smallest loss of $\mathcal{L}_{\max}$. On the other hand, sIB algorithm conducts sequential re-assignment of data objects, instead of the merger of clusters, so that the re-assignment leads to the smallest loss of $\mathcal{L}_{\max}$. Intuitively, these algorithms utilize a kind of distance measure based on $\mathcal{L}_{\max}$.

Among these algorithms, it is reported that sIB algorithm outperformed other algorithms in terms of the quality of clusters. This algorithm conducts hard assignment of data objects into clusters.

4 A graph-based approach

4.1 Preliminaries

A graph $G(\boldsymbol{V}, \boldsymbol{E})$ consists of a (finite) set of vertices $\boldsymbol{V}$, a set of edges $\boldsymbol{E}$ over $\boldsymbol{V} \times \boldsymbol{V}$. The set $\boldsymbol{E}$ can be interpreted as representing a binary relation on $\boldsymbol{V}$. A pair of vertices (v_i, v_j) is in the binary relation defined by a graph $G(\boldsymbol{V}, \boldsymbol{E})$ if and only if the pair $(v_i, v_j) \in \boldsymbol{E}$. Intuitively, each data item is mapped to a vertex in a graph G, and a pair

of vertices (v_i, v_j) is in $\boldsymbol{E}$ if and only if they are directly connected by an edge in the graph.

An edge-weighted graph $G(\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{W})$ is defined as a graph $G(\boldsymbol{V}, \boldsymbol{E})$ with the weight on each edge in $\boldsymbol{E}$. When $|\boldsymbol{V}| = n$, the weights in $\boldsymbol{W}$ can be represented as an n by n matrix $\mathbf{W}$, where w_{ij} in $\mathbf{W}$ stands for the weight on the edge for the pair $(v_i, v_j) \in \boldsymbol{E}$. In this paper, a bold capital letter $\mathbf{W}$ denotes a matrix representation of the set of weights. We set $w_{ij} = 0$ for pairs $(v_i, v_j) \notin \boldsymbol{E}$ so that the graph contain no self-loop.

4.2 A pseudo-similarity function

Based on Theorem 1 and (6), we regard that $D_{KL}[p(y|x)||p(y|t)]$ represents the pseudo-dissimilarity between x (data object) and t (cluster) under the framework in Section 3.[3] Furthermore, we extend this insight from $\mathcal{X} \times \mathcal{T}$ into $\mathcal{X} \times \mathcal{X}$, and propose to utilize KL-divergence as a pseudo-dissimilarity function between data objects for the clustering problem. Thus, $D_{KL}[p(y|x_i)||p(y|x_j)]$ represents the pseudo-dissimilarity among data objects x_i and x_j in our proposal.

Based on the above function, we also propose the following function under framework in Section 3. This function measures the relevance of pairwise objects and corresponds to a pseudo-similarity function.

Definition 4 s: $\mathcal{X} \times \mathcal{X} \to \mathcal{R}^+$ is a pseudo-similarity function, and defined as

$$s(x_i, x_j) = p(x_j) \exp(-\beta D_{KL}[p(y|x_i)||p(y|x_j)]) \tag{12}$$

where β is the parameter in Problem 1.

4.3 A data graph

The function in (12) represents the relation between each pair of objects. Since pairwise relation can be represented as a graph, we propose to represent them as an edge-weighted graph, where the values calculated by the function in (12) are utilized as their weights on the edges.

Definition 5 For a set of objects $\boldsymbol{X}$, by mapping each data object x to a vertex, an edge-weighted graph $G(\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{W})$ is defined as:

$$\boldsymbol{V} = \boldsymbol{X} \tag{13}$$

$$w_{ij} = \begin{cases} s(x_i, x_j) & x_i \neq x_j \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

From (13), we abuse the symbol $\boldsymbol{X}$ to denote the set of vertices in the data graph. Note that $w_{ij} \geq 0, \forall x_i, x_j \in \boldsymbol{X}$, in our definition. We call this graph the data graph in this paper. We assume that the data graph G for a given data objects $\boldsymbol{X}$ is connected.[4]

[3] Note that $D_{KL}[p(y|x)||p(y|t)]$ is not symmetric.

[4] Each vertex has at least one edge with positive weight. For disconnected graphs, each component can be dealt with separately.

Proposition 2 *$\frac{w_{ij}}{\sum_j w_{ij}}$ is a valid conditional probability distribution from x_i to x_j in the data graph.*

Proof By Definition 4 and (14), $w_{ij} \geq 0$ for all the weights. Since each data object x_i has at least one edge with positive weight, $\sum_j w_{ij} > 0$ and $0 \leq \frac{w_{ij}}{\sum_j w_{ij}} \leq 1$. By definition, $\sum_j \frac{w_{ij}}{\sum_j w_{ij}} = 1$. Thus, $\frac{w_{ij}}{\sum_j w_{ij}}$ satisfies the axioms of probability. □

Based on Proposition 2, we define the conditional probability in the data graph as

$$p(x_j|x_i) = \frac{w_{ij}}{\sum_j w_{ij}} \tag{15}$$

Here, $p(x_j|x_i)$ represents the conditional probability of seeing a data object x_j from the data object x_i. In other words, it corresponds to the transition probability from data object x_i to x_j in Markov model over the data graph.

Proposition 3 *The conditional probability in* (15) *is a stationary distribution in Theorem* 1 *where* $\boldsymbol{T} = \boldsymbol{X}$.

Proof By treating each data object x_j as an independent cluster t, it is easy to confirm from (12) and (14). □

As in hierarchical clustering methods (Jain et al. 1999; Guha et al. 1998), setting $\boldsymbol{T} = \boldsymbol{X}$ corresponds the situation where each data object is treated as an independent cluster. In that situation, Proposition 3 says that $p(x_j|x_i)$ in (15) over the data graph, based on the proposed function in (12), satisfies the necessary condition for the optimal solution of Problem 1.

4.4 A graph-based formalization

We shall show that, for hard assignment, Problem 1 can be approximated as the following problem over the data graph when data is uniformly distributed.

Problem 2 When the number of clusters k is specified, find k disjoint subsets $\{\boldsymbol{E}_1, \ldots, \boldsymbol{E}_k\}$ of edges in the data graph G which minimizes

$$J = \sum_{t=1}^{k} \sum_{w_{ij} \in \boldsymbol{E}_t} w_{ij} \tag{16}$$

and the removal of $\{\boldsymbol{E}_1, \ldots, \boldsymbol{E}_k\}$ from G results in k disconnected components.

Removing edges $\boldsymbol{E}_t$ from the data graph amounts to disconnecting the t-th component (cluster) from the graph, and the sum of weights on the removed edges corresponds to inter-cluster similarity between the cluster and its complement in the data graph. Thus, the objective function can be seen as a kind of inter-cluster similarity among clusters, and minimizing (16) corresponds to finding well-separated clusters.

We show the correspondence between the original problem (Problem 1) and the combinatorial problem over the data graph (Problem 2) in the following. First of all, we define the sum of weights on the edges from x_i as d_i.[5]

$$d_i = \sum_{x_j} w_{ij}, \quad \forall x_i \in \boldsymbol{X} \tag{17}$$

4.4.1 Objective functions

Note that when random variables X and Y are specified, $I(X;Y)$ is some constant value for the given data objects $\boldsymbol{X}$. Based on this fact, Problem 1 can be transformed into the following equivalent problem for any fixed β (see Tishby et al. 1999; Slonim 2002).

Problem 3 Find the conditional probability distribution $p(t|x)$, which minimizes the following objective function

$$F_{IB} = \sum_x \sum_t p(x)p(t|x)(-\log Z(x,\beta)) \tag{18}$$

where $Z(x,\beta)$ is the function defined in (7).

Proof (from Section 3.1.1 in Slonim (2002))

$$\begin{aligned}
\mathcal{L} + \beta I(X;Y) &= I(X;T) + \beta\{I(X;Y) - I(T;Y)\} \\
&= \sum_x \sum_t p(x)p(t|x) \log \frac{p(t|x)}{p(t)} \\
&\quad + \beta \sum_x \sum_t p(x)p(t|x) D_{KL}[p(y|x)||p(y|t)] \\
&= \sum_x \sum_t p(x)p(t|x)(-\log Z(x,\beta)) \\
&= F_{IB}
\end{aligned} \tag{19}$$

Thus, minimization of $\mathcal{L}$ is equivalent to the minimization of F_{IB} (modulo the constant $\beta\, I(X;Y)$).[6] □

The objective function in the data graph G, which corresponds to the one in (18), is represented as

$$F_G = \sum_{x_i} \sum_{x_j} p(x_i)p(x_j|x_i)(-\log Z(x_i,\beta)) \tag{20}$$

[5] $\sum_{x_j}$ ranges over $\boldsymbol{X}$ and corresponds to $\sum_j$.

[6] $I(X;Y) - I(T;Y) = \sum_{x,y} p(x,y) \log \frac{p(y|x)}{p(y)} - \sum_{y,t} p(y,t) \log \frac{p(y|t)}{p(y)} = \sum_{x,y,t} p(x,y,t) \left(\log \frac{p(y|x)}{p(y|t)} + \log \frac{p(y|t)}{p(y)}\right) - \sum_{x,y,t} p(x,y,t) \log \frac{p(y|t)}{p(y)} = \sum_x \sum_t p(x)p(t|x) \sum_y p(y|x) \log \frac{p(y|x)}{p(y|t)} = \sum_x \sum_t p(x)p(t|x) D_{KL}[p(y|x)||p(y|t)]$

We introduce one assumption to show our result.

Assumption 1 Data is uniformly distributed and $p(x)$ is some constant $c(=1/|\boldsymbol{X}|)>0$.

Hereafter, Assumption 1 is called uniform distribution, and assumed in the rest of the paper.

Proposition 4 *Under uniform distribution, F_G is some constant for a given data set $\boldsymbol{X}$.*

Proof

$$F_G = \sum_{x_i}\sum_{x_j} p(x_i)p(x_j|x_i)(-\log Z(x_i,\beta))$$

$$= c\sum_{x_i}(-\log Z(x_i,\beta))\sum_{x_j} p(x_j|x_i) \tag{21}$$

$$= c\sum_{x_i}(-\log d_i)\sum_{x_j}\frac{w_{ij}}{d_i} \tag{22}$$

$$= c\sum_{x_i}(-\log d_i) \tag{23}$$

Note that $p(x_i)=c$, $Z(x_i,\beta)=\sum_{x_j} p(x_j)exp(-\beta D_{KL}[p(y|x_i)||p(y|x_j)])=\sum_{x_j} w_{ij}=d_i$, and d_i is some constant for each x_i. Thus, (21) follows. Equation (22) follows from Proposition 2 and $\sum_{x_j}\frac{w_{ij}}{d_i}=1$ for each data x_i induces (23). Since each $-\log d_i$ is some constant as in (21), Proposition 4 holds. □

4.4.2 Compression and cut

Let us consider a 2-way partition of $\boldsymbol{X}$ into two mutually exclusive and exhaustive sets, i.e., $\boldsymbol{X}=S\sqcup\bar{S}$[7] (von Luxburg 2007). S and $\bar{S}$ corresponds to two clusters of objects. By removing or cutting the edges between S and $\bar{S}$, the data graph G is partitioned into two induced subgraphs G_S and $G_{\bar{S}}$ (Diestel 2006), and becomes a (disconnected) graph $\hat{G}=\{G_S, G_{\bar{S}}\}$.

Definition 6 We define the following to characterize a partition.

$$cut(S,\bar{S}) = \sum_{x_i\in S}\sum_{x_j\in\bar{S}} w_{ij} \tag{24}$$

$$cut(\bar{S},S) = \sum_{x_i\in\bar{S}}\sum_{x_j\in S} w_{ij} \tag{25}$$

[7] $\bar{S}$ is the complement of S. We follow the convention to utilize the symbol S to denote the subset in a partition.

Proposition 5 *For any partition of the data graph G where each induced subgraph G_S with $|S| > 1$, $\frac{w_{ij}}{\sum_{j \in G_S} w_{ij}}$ is a valid conditional probability distribution in G_S.*

Proof Since each G_S is an induced subgraph of G and $|S| > 1$, it is connected by definition. Thus, Proposition 5 follows from Proposition 2. □

For each $x_i \in \boldsymbol{X}$, let us denote S_i for the subset of $\boldsymbol{X}$ which contains x_i, and $\bar{S}_i$ for the other.

As in (17), we define the followings.

$$d_{S_i} = \sum_{x_j \in S} w_{ij} \tag{26}$$

$$d_{\bar{S}_i} = \sum_{x_j \in \bar{S}} w_{ij} \tag{27}$$

The following relation holds between (17) and (27) for each x_i in G.

$$d_i = d_{S_i} + d_{\bar{S}_i} \tag{28}$$

We define the conditional probability distribution in $\hat{G}$ as

$$\forall x_i \in S, \quad \hat{p}(x_j|x_i) = \begin{cases} \frac{w_{ij}}{d_{S_i}} & x_j \in S \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

$$\forall x_i \in \bar{S}, \quad \hat{p}(x_j|x_i) = \begin{cases} \frac{w_{ij}}{d_{S_i}} & x_j \in \bar{S} \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

The induced subgraphs G_S and $G_{\bar{S}}$ are connected as in G, and the conditional probability defined in (29) and (30) are valid conditional probability in each subgraphs. Thus, as in (20), F_{G_S} and $F_{G_{\bar{S}}}$ are defined as:

$$F_{G_S} = \sum_{x_i \in S} \sum_{x_j \in S} p(x_i) \frac{w_{ij}}{d_{S_i}} (-\log Z(x_i, \beta)) \tag{31}$$

$$F_{G_{\bar{S}}} = \sum_{x_i \in \bar{S}} \sum_{x_j \in \bar{S}} p(x_i) \frac{w_{ij}}{d_{S_i}} (-\log Z(x_i, \beta)) \tag{32}$$

Although the partitioned graph $\hat{G}$ is disconnected, the objective function over $\hat{G}$ is defined as

$$F_{\hat{G}} = \sum_{x_i} \sum_{x_j} p(x_i) \hat{p}(x_j|x_i) (-\log Z(x_i, \beta)) \tag{33}$$

Under uniform distribution, $F_{\hat{G}}$ can be divided into the objective functions over the induced subgraphs, as follows:

$$F_{\hat{G}} = \sum_{x_i} \sum_{x_j} p(x_i)\hat{p}(x_j|x_i)(-\log Z(x_i, \beta)) \tag{34}$$

$$= \sum_{x_i \in S} \sum_{x_j \in S} p(x_i)\frac{w_{ij}}{d_{S_i}}(-\log Z(x_i, \beta)) + \sum_{x_i \in \bar{S}} \sum_{x_j \in \bar{S}} p(x_i)\frac{w_{ij}}{d_{S_i}}(-\log Z(x_i, \beta)) \tag{35}$$

$$= F_{G_S} + F_{G_{\bar{S}}} \tag{36}$$

$$= c\sum_{x_i \in S}(-\log d_{S_i}) + c\sum_{x_j \in \bar{S}}(-\log d_{S_j}) \tag{37}$$

Equation (37) can be derived from (36) as in Proposition 4.

Note that $\hat{p}(x_j|x_i)$ in (29) and (30) does not satisfy (6), since $p(S_i|x_i) = 1$ and $p(\bar{S}_i|x_i) = 0$ for all $x_i \in \boldsymbol{X}$,[8] and deviates from (6) due to the hard assignment of each x_i into S_i.[9] We would like to minimize the deviation to solve Problem 1. From Proposition 4, F_G is some constant for the given data $\boldsymbol{X}$. Thus, minimization of the deviation $F_{\hat{G}} - F_G$ is equivalent to the following problem.

Problem 4 For the data graph G, find the 2-way partition $\boldsymbol{X} = S \sqcup \bar{S}$ which minimizes $F_{\hat{G}} = F_{G_S} + F_{G_{\bar{S}}}$ in $\hat{G} = \{G_S, G_{\bar{S}}\}$.

4.4.3 Main result

Finally, we shall show our main result. First, we define the following problem.

Problem 5 For the data graph G, find two disjoint subsets $\{\boldsymbol{E}_1, \boldsymbol{E}_2\}$ of edges which minimize

$$J = \sum_{t=1}^{2} \sum_{w_{ij} \in \boldsymbol{E}_t} w_{ij} \tag{38}$$

and the removal of $\{\boldsymbol{E}_1, \boldsymbol{E}_2\}$ from G results in a disconnected graph $\hat{G} = \{G_S, G_{\bar{S}}\}$ where G_S and $G_{\bar{S}}$ are components of $\hat{G}$.

Claim In hard assignment, Problem 1 can be approximated to Problem 5 under uniform distribution.

[8] S and $\bar{S}$ corresponds to clusters.

[9] Any hard assignment deviates from (6).

Proof As described above, Problem 1 can be reduced to Problem 4. Thus, we approximate Problem 4 as Problem 5. In the following, symbol $\Leftrightarrow$ represents the equivalence, and symbol $\simeq$ represents the approximation.

$$\min F_{\hat{G}} \Leftrightarrow \min \left\{ \sum_{x_i \in S} (-\log d_{S_i}) + \sum_{x_j \in \bar{S}} (-\log d_{S_j}) \right\}$$

$$\simeq \min \left\{ \sum_{x_i \in S} d_{\bar{S}_i} + \sum_{x_j \in \bar{S}} d_{\bar{S}_j} + \sum_{x_i \in S \sqcup \bar{S}} (1 - d_i) \right\} \tag{39}$$

$$\Leftrightarrow \min \left\{ \sum_{x_i \in S} d_{\bar{S}_i} + \sum_{x_j \in \bar{S}} d_{\bar{S}_j} \right\} \tag{40}$$

$$\Leftrightarrow \min\{cut(S, \bar{S}) + cut(\bar{S}, S)\} \tag{41}$$

$$\Leftrightarrow \min \sum_{t=1}^{2} \sum_{w_{ij} \in \boldsymbol{E}_t} w_{ij} = \min J \tag{42}$$

From (37), the first equation holds under uniform distribution. Based on (28), by approximating the log function via Taylor expansion as $(-\log d_{S_i}) \simeq d_{\bar{S}_i} + (1 - d_i)$, (39) holds. As in Proposition 4, since each d_i is some constant in G, we can omit the third term in (39), and thus (39) is equivalent to (40). From (24) and (27), by the definition of $cut(S, \bar{S})$, (40) is equivalent to (41). Note that $cut(S, \bar{S})$ corresponds to the sum of weights on the edges from cluster S to $\bar{S}$ (and $cut(\bar{S}, S)$ is the other direction). Thus, $cut(S, \bar{S}) + cut(\bar{S}, S)$ in (41) corresponds to inter-cluster similarity between cluster S and $\bar{S}$. Also, $cut(S, \bar{S}) + cut(\bar{S}, S)$ amounts to $\sum_{t=1}^{2} \sum_{w_{ij} \in \boldsymbol{E}_t} w_{ij}$, and the latter is equivalent to (38). Thus, the *Claim* holds. □

The above *Claim* can be easily extended to the following general k-way partition where the number of clusters k is specified.

Claim In hard assignment, Problem 1 can be approximated to Problem 2 under uniform distribution.

4.5 Clustering based on data graph

From the above result, the clustering problem in Section 3 can be tackled by solving the combinatorial problem (Problem 2) over the proposed data graph. Various graph algorithms have been proposed for solving this kind of problem efficiently (Stoer and Wagner 1997). Such algorithms can be utilized to solve the clustering problem based on mutual information over the data graph.

However, it is known that small unbalanced clusters tend to be created under the minimum cut formulation of partitioning (von Luxburg 2007). From the objective of data clustering, unbalanced clusters are not desirable. Thus, when solving Problem 2 over the data graph, in addition to minimizing the objective function, it would be important to consider the balance between clusters.

Table 1 Datasets from 20 Newsgroup dataset

Dataset	Included group
Multi5	comp.graphics, rec.motorcycles,rec.sport.baseball,sci.space,talk.politics.mideast
Multi10	alt.atheism, comp.sys.mac.hardware,misc.forsale, rec.autos,rec.sport.hockey, sci.crypt,sci.med, sci.electronics,sci.space,talk.politics.guns
Multi15	alt.atheism, comp.graphics, comp.sys.mac.hardware, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.guns, talk.politics.mideast, talk.politics.misc

5 Evaluations

5.1 Application for text clustering

Based on previous work (Slonim 2002; Dhillon et al. 2003), we evaluated the proposed approach on the text clustering problem. As shown in the example in Section 3.2, for a given documents $\boldsymbol{X}$, the set of terms which are utilized to describe the documents correspond to $\boldsymbol{Y} = \{y_1, \ldots, y_m\}$, and $p(x, y)$ corresponds to the joint probability of a document x and a term y. The number of terms are huge in general. Thus, this problem corresponds to the clustering of high-dimensional sparse data, a rather difficult clustering problem. Since the proposed approach is a partitioning based method, we assume that the number of clusters k is specified.

As in Slonim (2002) and Dhillon et al. (2003), we evaluated the proposed approach over 20 Newsgroup data (20NG),[10] which is widely utilized as a benchmark in text processing community. We selected 3 sets of groups, which are called Multi5, Multi10 and Multi15 respectively, as shown in Table 1. We sampled 50 documents from each group in order to create one sample for each dataset. We repeated this process and created 10 samples for each set of groups. For each sample, we conducted stemming using porter stemmer[11] and MontyTagger,[12] removed stop words, and selected 2,000 words with large mutual information (Cover and Thomas 2006).

We also conducted evaluations over TREC datasets.[13] The characteristics of the utilized datasets are summarized in Table 2. We followed the same procedure in 20NG and created ten samples for each dataset. Since these datasets are already preprocessed and represented as count data, we did not conduct stemming or tagging.

5.2 Experimental settings

5.2.1 Spectral clustering over the data graph

For each dataset, we constructed the data graph in Section 4.3 and conducted clustering by solving Problem 2 over the graph. As described in Section 4.5, it is

[10] http://people.csail.mit.edu/jrennie/20Newsgroups/20news-18828 was utilized.

[11] http://www.tartarus.org/~martin/PorterStemmer

[12] http://web.media.mit.edu/~hugo/montytagger

[13] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download

Table 2 Characteristics of the original TREC datasets

Dataset	# Attr.	# Classes	# Data
Hitech	126,373	6	2,301
Reviews	126,373	5	4,069
New3	83,487	44	9,558

important to consider the balance among clusters. We utilized spectral clustering to fulfill this requirement (von Luxburg 2007).

The goal of clustering is to assign similar objects to the same cluster while dissimilar ones to different clusters. In spectral clustering (von Luxburg 2007), this is realized by seeking a function $f\colon \mathcal{X} \to \mathcal{R}$, which assigns similar values for similar objects and dissimilar values for dissimilar objects. By assuming that the pairwise similarities among data objects can be specified, the data objects $\boldsymbol{X}$ is represented as an edge-weighted graph, where each pair of vertices are connected with an edge with their similarity. The objective function is represented as

$$J_1 = \sum_{i,j} w_{ij}(f_i - f_j)^2 \tag{43}$$

where i, j sum over the vertices in the graph, w_{ij} corresponds to the similarity between data objects x_i and x_j, $f_i = f(x_i)$ for each x_i in $\boldsymbol{X}$. Thus, data clustering is formalized as an optimization problem to find out the function f which minimizes the objective function.

Formally, the similarities can be represented as a symmetric[14] n-by-n square matrix $\mathbf{W}$ when $|\boldsymbol{X}| = n$. Based on the weights in the graph, the objective function in (43) can be reduced to finding the generalized eigenvector of the following matrix $\mathbf{L}$, which is called graph Laplacian in spectral graph theory (Chung 1997):

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{44}$$

$$\mathbf{D} = diag(d_1, \ldots, d_n) \tag{45}$$

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{46}$$

$$\mathbf{L}\boldsymbol{h} = \lambda \mathbf{D}\boldsymbol{h} \tag{47}$$

where d_i in (44) corresponds to (17) in our data graph, and λ is the generalized eigenvalue. The generalized eigenvector $\boldsymbol{h}$ of $\mathbf{L}$ in (47) is considered as the embedded representation of the data set $\boldsymbol{X}$. For a specified number l, l eigenvectors $\mathbf{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_l\}$ with the smallest non-zero eigenvalues are constructed. These correspond to spectral embedding of $\boldsymbol{X}$ onto the subspace spanned by $\mathbf{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_l\}$ (Belkin and Niyogi 2002).[15] Some clustering method is applied to $\mathbf{H}$ and the constructed clusters are returned.

[14] Although it is possible to deal with asymmetric matrix, we focus on symmetric one in this paper.

[15] l corresponds to the number of dimension of the embedded subspace.

In order to balance the clusters, two representative normalized graph Laplacian have been proposed based on $\mathbf{D}$ in (45) (von Luxburg 2007):

$$\mathbf{L}_{rw} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W} \tag{48}$$

$$\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}} \tag{49}$$

In the above two normalized graph Laplacian, $\mathbf{L}_{rw}$ is based on the random walk over the graph, and $\mathbf{L}_{\text{sym}}$ is based on the symmetric normalization. Both of them have been widely utilized in the literature.

5.2.2 Compared methods

As for the proposed approach, we evaluated the clustering over the data graph with $\mathbf{L}_{rw}$ and $\mathbf{L}_{\text{sym}}$, respectively. The constructed embedded representation over the data graph is represented as $\mathbf{H}_{rw}$, and the one with $\mathbf{L}_{\text{sym}}$ as $\mathbf{H}_{\text{sym}}$. For each pair (x_i, x_j) the edges with w_{ij} and w_{ji} are defined in the data graph. However, these should be removed *simultaneously* for partitioning, as defined in Problem 2. Thus, we set the symmetric matrix $(\mathbf{W})_{ij} = (w_{ij} + w_{ji})/2$ in the following experiment. Clustering was conducted on each embedded representation using the spherical kmeans (skmeans) algorithm.

We compared the proposed approach with iIB and sIB in Tishby et al. (1999) and Slonim et al. (2002), and with spherical kmeans (skmeans) (Dhillon and Modha 2001). Spherical k-means algorithm (skmeans) was proposed for large sparse text data clustering. It is an extension of the standard kmeans algorithm. The difference is that cosine similarity is utilized as a similarity measure in skmeans, and each data is re-assigned to the "nearest" cluster in terms of cosine similarity. As described in Section 3.3, iIB tries to find the stationary distribution in (6) via projection, and sIB conducts sequential re-assignment of data into clusters. The joint probability $p(x, y)$ was estimated from each dataset using Ristad method (Ristad 1995).

Ristad method (Ristad 1995) is a discounting method to cope with the zero frequency problem in natural language processing. It is based on the properties of naturally occurring string, and calculates the probability of string. In our setting, a string corresponds to a term. Suppose we are given a text corpus. N stands for the length of the corpus (i.e., the total number of terms), and V stands for the number of different possible terms. For each term w, $C(w)$ stands for the number of occurrence of w in the corpus. N_0 stands for the number of terms which do not occur (i.e., zero-frequency terms) in the corpus. The probability $P(w)$ is calculated in the Ristad method as follows.

$$P(w) = \begin{cases} \dfrac{C(w)+1}{N+V} & N_0 = 0 \\ \dfrac{(C(w)+1)(N+1+N_0-V)}{N^2+N+2(V-N_0)} & N_0 > 0 \text{ and } C(w) > 0 \\ \dfrac{(V-N_0)(V-N_0+1)}{N_0(N^2+N+2(V-N_0))} & \text{otherwise} \end{cases} \tag{50}$$

5.2.3 Parameters

The main parameter in the framework of mutual information based clustering is the parameter β, and this is also utilized in (12). sIB makes it irrelevant to β by setting

it a very large value (β=10^4) (Slonim 2002); however, both IIB and the proposed approach are affected by the value of β. Thus, we conducted preliminary experiments and set the value as $\beta \in [15, 70]$ for IIB and as $\beta \in [10^{-3}, 1]$ for the proposed approach in the following experiments.

The number of dimension l (the number of eigenvectors) also affects the performance in spectral clustering. Basically it was set as $l = k$ (the number of clusters); however, for the datasets with less than 10 clusters, l was set to $k \times 2$ since a subspace with $l = k$ seems inappropriate (too low dimension).

5.3 Evaluation measures

For each dataset, cluster assignment was evaluated with respect to both external measures and internal measures. Note that both 20NG and TREC datasets contain the true cluster label for each data. External measures are calculated based on the "ground truth" labels in each dataset and constructed clusters. On the other hand, internal measures (also called as clustering validation indices) (Halkidi et al. 2002; Maulik and Bandyopadhyay 2002) are calculated in the original data space (i.e., with the original data representation) based on the cluster assignment.

5.3.1 External measures

Among various external measures (Strehl and Ghosh 2002), we evaluated Normalized Mutual Information (NMI) and Purity in the following experiment.

Let T, $\hat{T}$ stand for the random variables over the ground truth clusters and constructed clusters. NMI is defined as

$$NMI = \frac{I(\hat{T}; T)}{(H(\hat{T}) + H(T))/2} \quad (\in [0, 1]) \tag{51}$$

where $H(T)$ is Shannon Entropy. The larger NMI is, the more correct the cluster assignment is with respect to the true cluster labels.

Purity is defined based on the contingency table of the ground truth cluster G_i and the constructed cluster C_h as follows:

$$Purity = \frac{1}{n} \sum_{i=1}^{k} \max_h |G_i \cap C_h| \quad (\in [0, 1]) \tag{52}$$

where n is the number of data. As shown in (52), it is calculated based on the number of shared data objects among G_i and C_h. The larger Purity is, the more cohesive the constructed clusters are.

5.3.2 Internal measures

Among various internal measures (Halkidi et al. 2002; Maulik and Bandyopadhyay 2002), we evaluated (i) Dunn's index (Dunn), (ii) Davis–Bouldin's index (D.B.), (iii) average scattering for clusters (Scat), (iv) total scattering (separation) between clusters (Dis). Although Scat and Dis are two terms utilized in SD validity index (Halkidi et al. 2002), a weighting factor needs to be specified to calculate SD validity index (since these are of the different range). However, since the weighting factor also affects the value of the index, we evaluated them separately.

In the following, internal measures were calculated in the data space (i.e., with respect to the given representation of the data). $\mathbf{X}$ stands for a sample in the data space, $\boldsymbol{c_i}$ stands for the center of cluster $C_i (i = 1 \ldots k)$. For a vector $\boldsymbol{x}$, $||\boldsymbol{x}||$ is defined as $||\boldsymbol{x}|| = (\boldsymbol{x}^T \boldsymbol{x})^{1/2}$. For the entire data $\mathbf{X}$, $\sigma(\mathbf{X})$ stands for a vector of variance in each dimension, and $\sigma(C_i)$ for that of cluster C_i.

For clusters C_i and C_j, let $diam(C_i)$ stands for the diameter of the cluster C_i, $dist(C_i, C_j)$ for the distance (dispersion) between clusters C_i and C_j. Euclidian distance was utilized as the distance metric $d(\cdot, \cdot)$ in the following evaluation.

(i) Dunn's Index (Dunn):

$$Dunn = \min_{i=1\ldots k} \min_{j=i+1\ldots k} \frac{dist(C_i, C_j)}{\max_{h=1\ldots k} diam(C_h)} \tag{53}$$

Clusters with larger Dunn's index are considered as better ones.

(ii) Davis–Bouldin's index (D.B.):

$$D.B. = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{diam(C_i) + diam(C_j)}{dist(C_i, C_j)} \tag{54}$$

Clusters with smaller D.B. index are considered as better ones.
In the original definition (Halkidi et al. 2002), $diam(\cdot)$ is defined as the complete diameter, i.e., $diam(C_i) = \max_{\boldsymbol{x}, \boldsymbol{y} \in C_i} d(\boldsymbol{x}, \boldsymbol{y})$, and $dist(\cdot, \cdot)$ is defined as the single linkage distance, i.e., $dist(C_i, C_j) = \min_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})$. However, as in density-based clustering methods (Ester et al. 1996), since no assumption about the shape of clusters is made in our approach, it is also possible to consider them via *average* distance, i.e., $diam(C_i) = \frac{1}{|C_i|(|C_i|-1)} \sum_{\boldsymbol{x}, \boldsymbol{y} \in C_i, \boldsymbol{x} \neq \boldsymbol{y}} d(\boldsymbol{x}, \boldsymbol{y})$ and $dist(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})$. We represent the indices with the complete diameter and single linkage distance as Dunn and D.B., and the average version of both indices as Dunn(ave) and D.B.(ave), respectively.

(iii) average scattering for clusters (Scat):

$$Scat = \frac{1}{k} \sum_{i=1}^{k} \frac{||\sigma(C_i)||}{||\sigma(\mathbf{X})||} \tag{55}$$

Scat corresponds to intra-cluster scatter. Thus, the smaller Scat is, the better the clusters are.

(iv) total scattering (separation) between clusters (Dis):

$$Dis = \frac{D_{\max}}{D_{\min}} \sum_{i=1}^{k} \left(\sum_{j=1}^{k} ||\boldsymbol{c}_i - \boldsymbol{c}_j|| \right)^{-1} \tag{56}$$

where $D_{\max} = \max_{i,j=1\ldots k} ||\boldsymbol{c}_i - \boldsymbol{c}_j||$, $D_{\min} = \min_{i,j=1\ldots k} ||\boldsymbol{c}_i - \boldsymbol{c}_j||$. Since Dis corresponds to inter-cluster separation, the larger Dis is, the better the clusters are.

5.4 Results

For each sample we conducted ten runs of experiment in order to account for the influence of initial configuration in clustering. Since ten samples were constructed for each dataset, the average of 100 runs were calculated for each dataset.

In the reported figures, kl-rw (red line) stands for the proposed approach with $\mathbf{L}_{rw}$, and kl-sym (purple line) for the proposed one with $\mathbf{L}_{\text{sym}}$. The compared methods are: sIB (blue line), iIB (green line), and skmeans (black dotted line)). Since β is the main parameter, x axis is for β and y axis is for the evaluation measure in the figures.

5.4.1 Results on 20NG

The results on 20NG with respect to external measures are shown in Fig. 2 (NMI) and Fig. 3 (Purity).

As for NMI (Fig. 2), which corresponds to the correctness of cluster assignment, the proposed method with $\mathbf{L}_{rw}$ (kl-rw) outperformed other methods (i.e., with larger NMI) with respect to Multi10 and Multi15. On the other hand, for Multi5, although it outperformed iIB and skmeans, but it was below sIB.

Furthermore, the proposed approach (both kl-rw and kl-sym) is quite stable for different values of β. Thus, the proposed approach can be considered as robust to

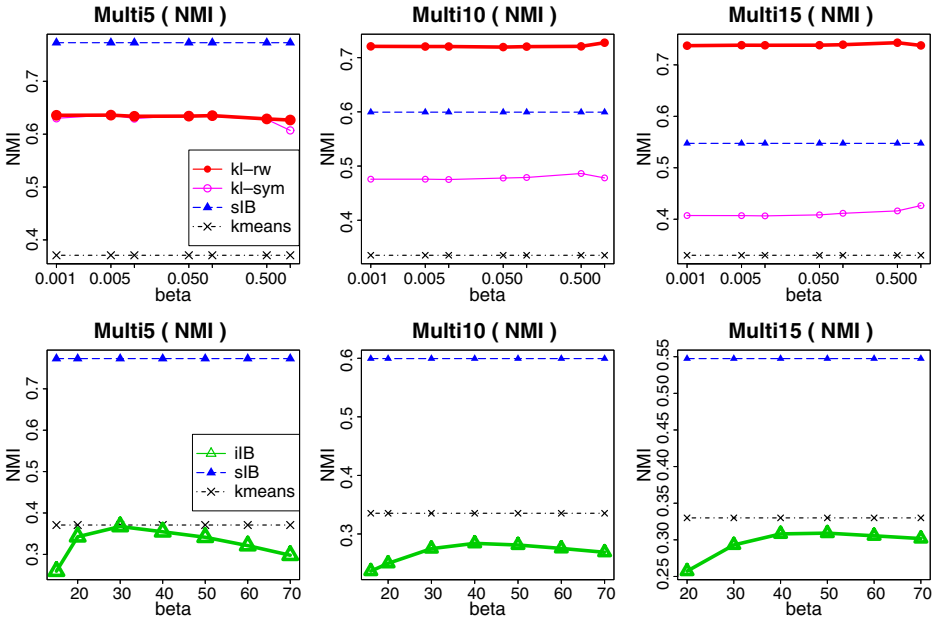

Fig. 2 Results on 20NG (w.r.t. NMI)

this parameter. On the other hand, in ilB, the performance varied depending on the value of β, but maximum NMI was obtained between 30 and 40.

As for Purity (Fig. 3), the results are similar to the results for NMI, and the proposed method with $\mathbf{L}_{rw}$ (kl-rw) outperformed other methods (with larger Purity) for Multi10 and Multi15, but not for Multi5.

The results with respect to internal measures are shown in Fig. 4 (Dunn and D.B.) and Fig. 5 (Scat and Dis). In terms of Dunn's index (the first column in Fig. 4), contrary to the external measures, the proposed approach could not outperform other methods. However, ilB, which was inferior to other methods with respect to NMI and Purity, showed the best performance (with larger Dunn).

On the other hand, in terms of Dunn(ave) with average diameter and cluster distance (the second column in Fig. 4), kl-rw outperformed other methods in all datasets. Since the objective function in (16) is defined as the sum of weights over the edges between clusters, indices based on the average diameter and distance seems reasonable. Furthermore, the proposed approach was stable with respect to β for Dunn(ave) (and also D.B.(ave)), as in the external measures.

For D.B. and D.B.(ave) indices (the third and fourth columns in Fig. 4), ilB with small β outperformed other methods (with smaller D.B.). Except for ilB, kl-sym was the best in all datasets.

For Scat, slB showed the best performance (with smaller Scat). This would be because slB works as a sequential kmeans algorithm with respect to Jensen–Shannon divergence (Cover and Thomas 2006) and constructs convex clusters (as in Voronoi partition). On the other hand, ilB was the worst, since it conducts iterative projection

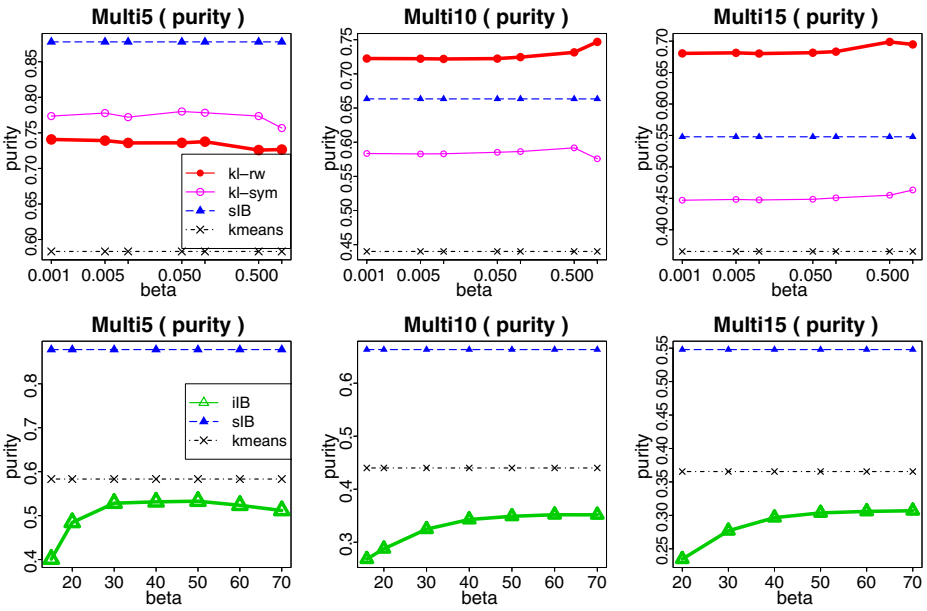

Fig. 3 Results on 20NG (w.r.t. Purity)

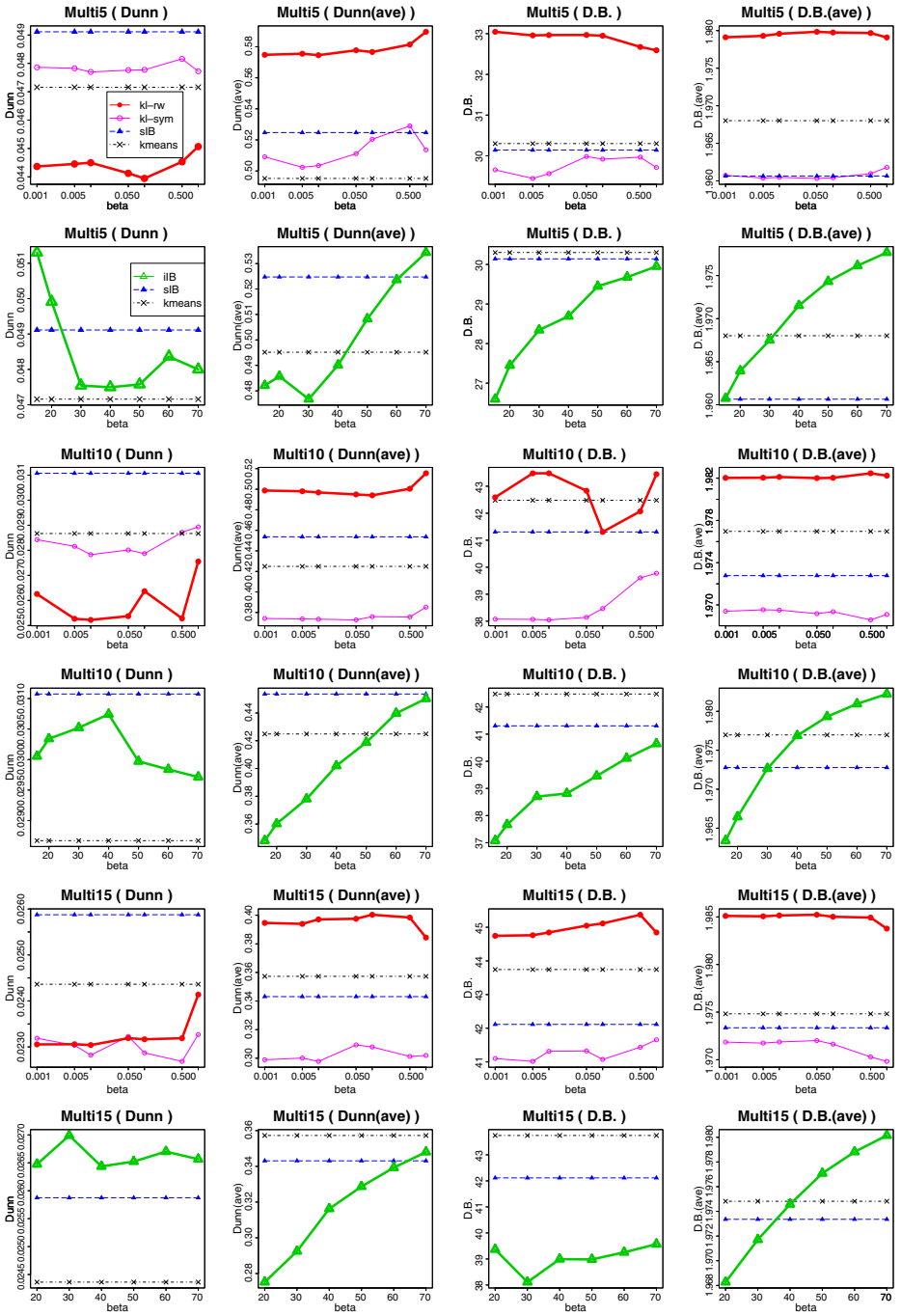

Fig. 4 Results on 20NG (Dunn and D.B.)

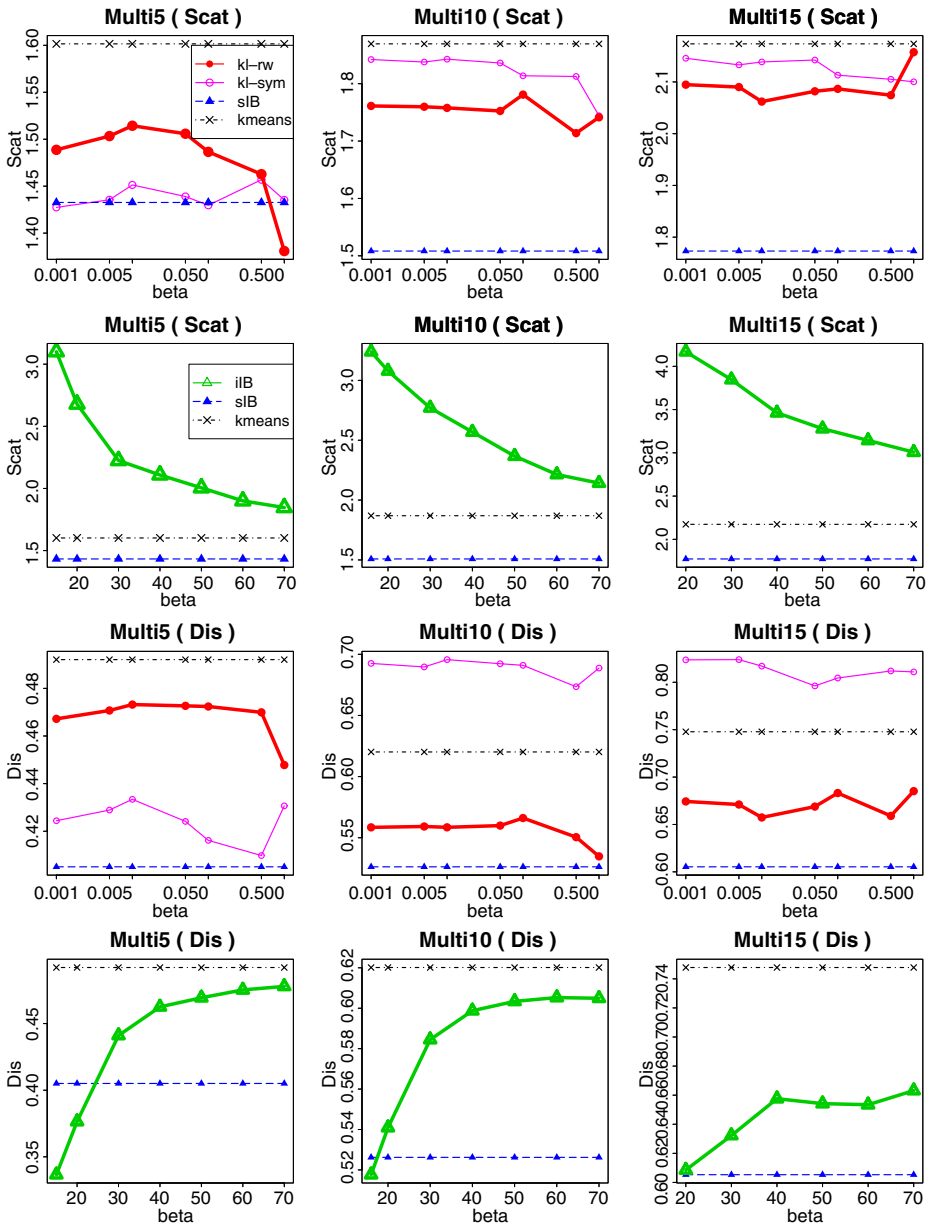

Fig. 5 Results on 20NG (Scat and Dis)

and does not consider the shape of clusters. Although our methods (kl-rw and kl-sym) also do not explicity consider the shape of clusters, it outperformed skmeans, which constructs convex clusters with respect to cosine similarity.

For Dis, our methods (kl-rw and kl-sym) outperformed sIB (with larger Dis). For Multi10 and Multi15, kl-sym showed the best performance.

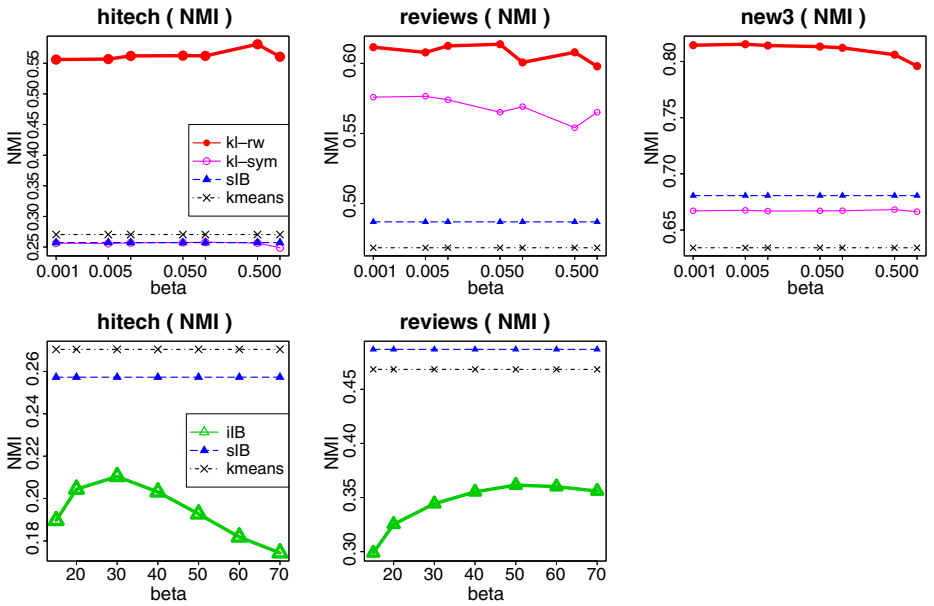

Fig. 6 Results on TREC (w.r.t. NMI)

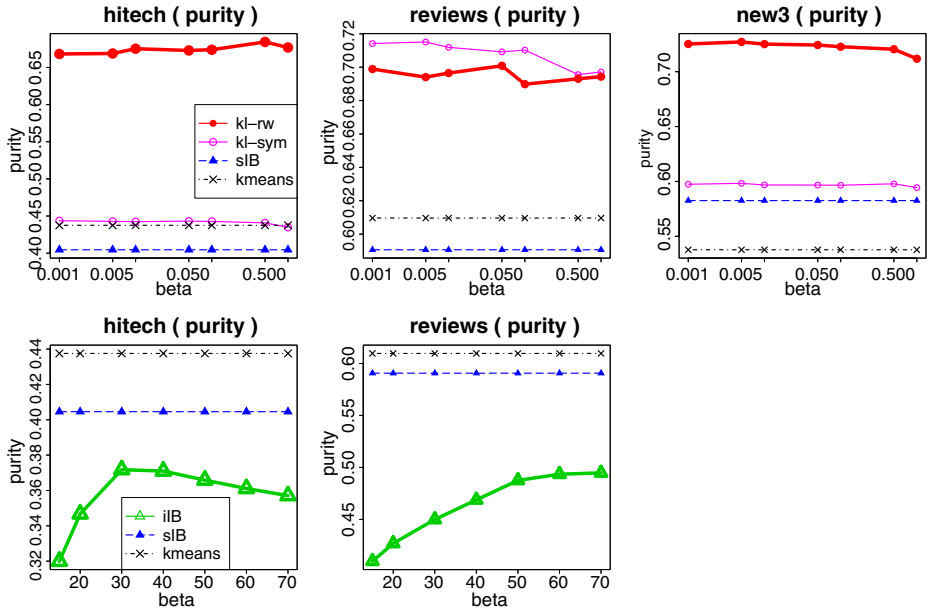

Fig. 7 Results on TREC (w.r.t. purity)

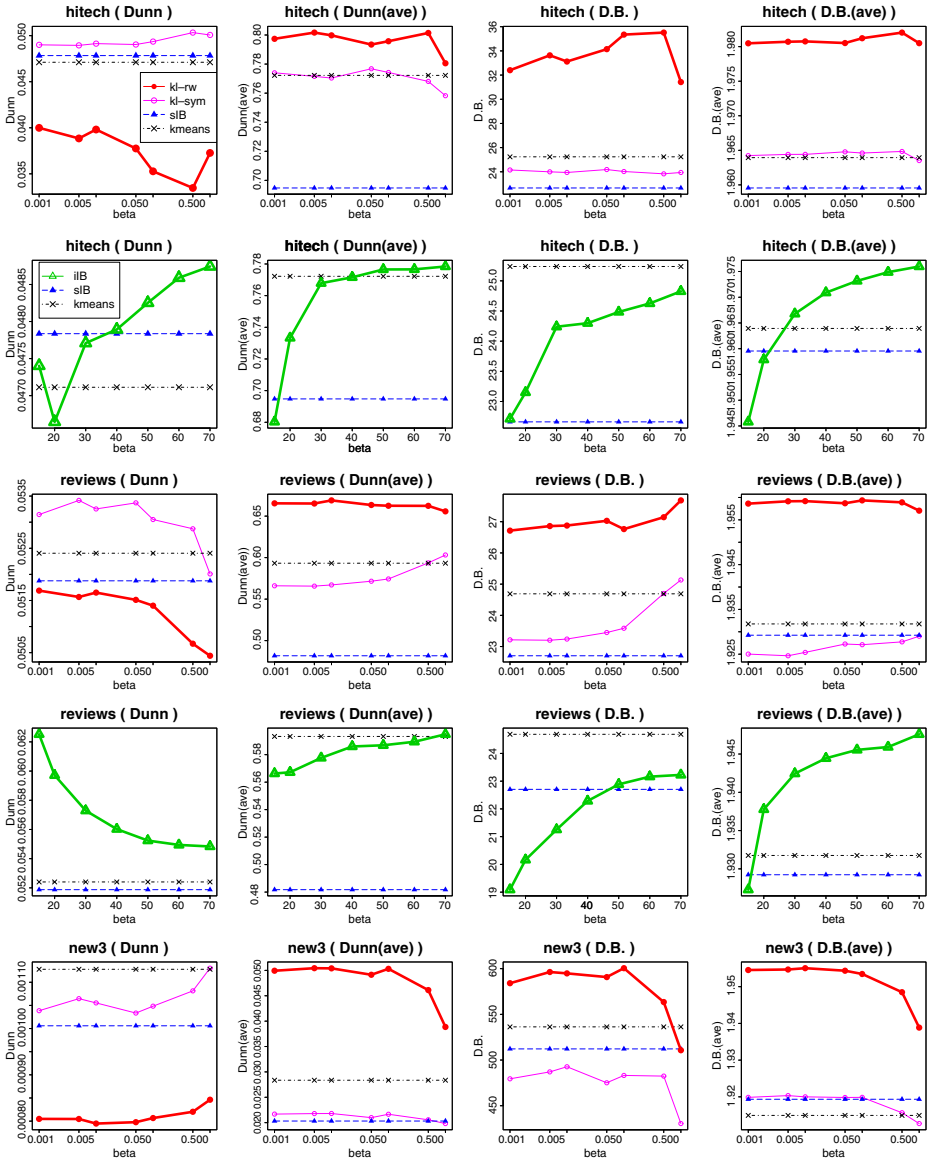

Fig. 8 Results on TREC (Dunn and D.B.)

5.4.2 Results on TREC

The results on TREC datasets are shown in Figs. 6 (NMI), 7 (Purity), 8 (Dunn and D.B.) and 9 (Scat and Dis). Since ilB takes too much time as the size of dataset increases, its results for new3 are not reported.[16]

[16] A dataset for new3 contains 2,200 data items. One run of ilB took more than 3 h, and we could not evaluate 100 runs for each value of β.

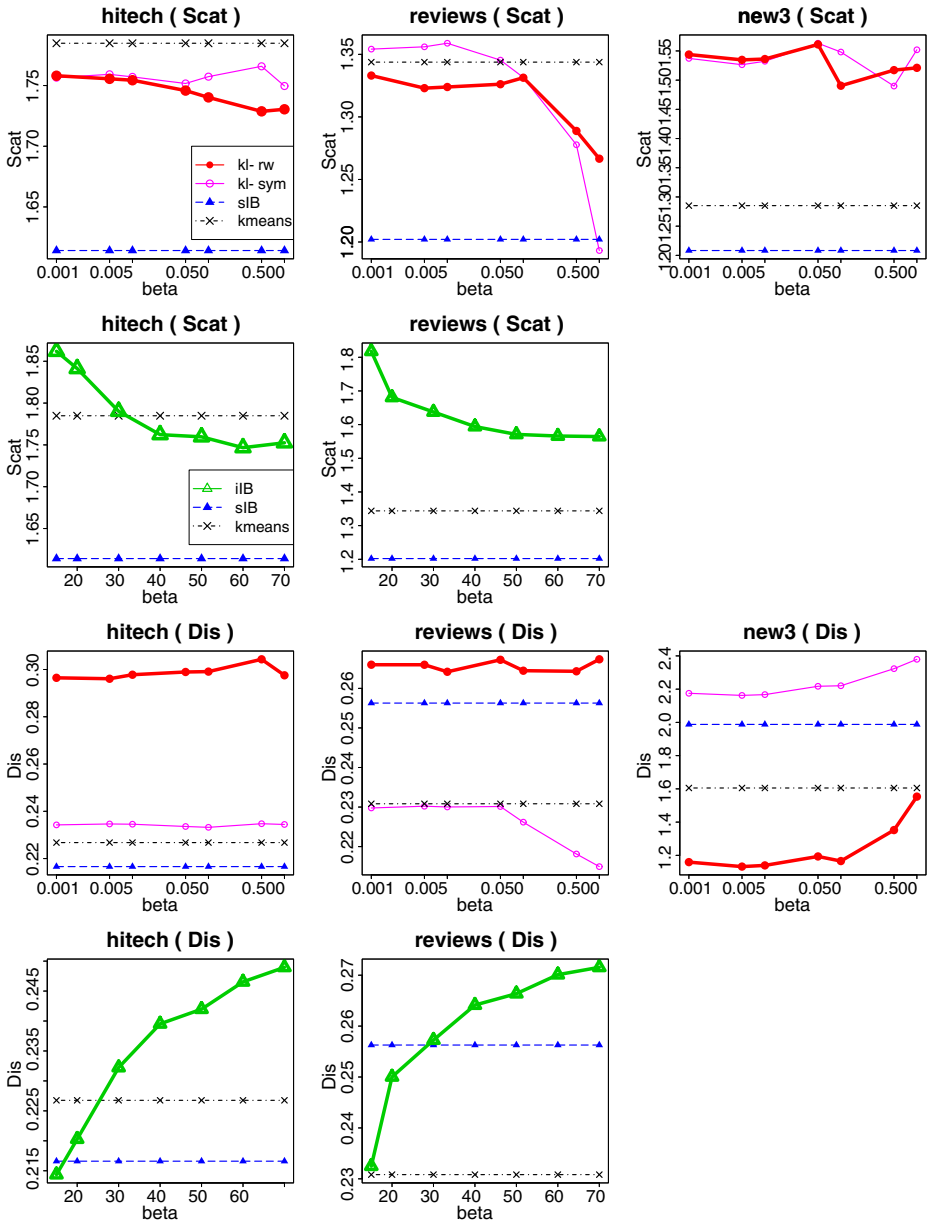

Fig. 9 Results on TREC (Scat and Dis)

On the whole, the results on TREC datasets were similar to the ones on 20NG. Our methods (especially kl-rw) outperformed others with respect to external measures (NMI and Purity). kl-sym showed the best performance for Dunn in hitech and reviews datasets. On the other hand, kl-rw showed the best performance for Dunn(ave) in all datasets, as in 20NG. iIB with small β was the best for both D.B.

and D.B.(ave), and slB was the best for Scat. However, for Dis, which corresponds to inter-cluster separation, kl-rw showed the best performance in hitech and reviews datasets, but not in new3.

5.5 Discussion

As described in Section 3.3, ilB algorithm utilized iterative projection based on the stationary distribution in Theorem 1. On the other hand, our approach is based on the approximation of the stationary distribution via cut. Thus, with respect to finding out the approximated solution based on the stationary distribution, the proposed approach corresponds to ilB algorithm. Since the proposed approach outperformed ilB in all the datasets with both NMI (51) and Purity (52), these results confirmed the validity and the effectiveness of the proposed approach with respect to external measures.

The proposed approach formalizes Problem 1 as the corresponding combinatorial problem based on the induced conditional probability over the data graph. $\mathbf{L}_{rw}$ conducts the normalization of graph Laplacian based on the random walk over the graph, which is induced from the weights of the graph (von Luxburg 2007). Thus, although both $\mathbf{L}_{rw}$ and $\mathbf{L}_{\mathrm{sym}}$ are widely utilized, the former seems to match the proposed approach in terms of the conditional probability interpretation. In addition, the experimental results also validate that $\mathbf{L}_{rw}$ is more suitable for the data graph. Thus, the proposed approach can be considered as a valid model for data clustering based on mutual information in Section 3, in terms of both the interpretation and the experimental evaluation.

We applied the proposed approach for text clustering, which is a rather difficult problem due to high-dimensionality and sparseness of the representation. Since the proposed method (kl-rw) outperformed slB for the datasets except Multi5, these results showed the effectiveness of the proposed approach. One of the reasons for the result in Multi5 is that, Problem 1 is formalized based on KL divergence in (1). However, it can be rather numerically instable when the zero frequency problem in text processing occurs. Coping with this problem is left for future research work.

Various cluster validation indices have been proposed and utilized (Ghosh 2003; Halkidi et al. 2002). Since NMI corresponds to the correctness of assignment of data with respect to the "ground truth" of the assignment, it reflects more the quality of constructed clusters than Purity. Proposed methods (both kl-rw and kl-sym) outperformed other methods in most datasets for NMI. Especially, kl-rw was the best. Thus, the proposed approach can be said as effective and has appealing clustering performance with respect to externally predefined labels.

For internal measures such as Dunn (53) and D.B. (54), our approach could not outperform other methods. However, with respect to average diameter and cluster distance, kl-rw showed the best performance for Dunn(ave) in all datasets. For Scat (55), although our approach does not explicitly try to construct convex clusters, still it outperformed skmeans (but not slB). On the other hand, for Dis (56), which corresponds to inter-cluster separation, proposed approach showed good performance. This result would be partly because the objective function in (16) corresponds to inter-cluster similarity and is minimized by our approach.

As in density-based clustering methods (Ester et al. 1996), no assumption about the shape of clusters is made in our approach. As illustrated in Roweis and Saul

(2000) and Tenenbaum et al. (2000), it is known that "swiss roll" shaped clusters are difficult to identify in the data space. Furthermore, such clusters would be evaluated as low quality in terms of internal measures (especially (53) and (54)) in the data space, since cluster diameter $diam(\cdot)$ is large and cluster distance $dist(\cdot, \cdot)$ is relatively small in swiss roll shaped clusters. Thus, although our approach did not always outperform other methods in terms of internal measures in the data space, since it showed a good performance in terms of external measures, we believe that it is worthwhile to pursue the proposed approach.

6 Concluding remarks

We proposed a graph model for the clustering problem based on mutual information (Tishby et al. 1999). We have shown that, in hard assignment where each object is assigned only to one cluster, the clustering problem can be approximated as a combinatorial problem over the proposed edge-weighted graph when data is uniformly distributed. Based on the stationary distribution induced from mutual information, we proposed a function which measures the relevance among data objects under the problem setting, and utilized it to map the entire data objects into the corresponding graph. The proposed graph model enables to transform the original clustering problem to a combinatorial problem over the graph.

Representing the entire data objects as a graph based on our graph model and formalizing the clustering problem over the graph enable to utilize various graph algorithms to solve the clustering problem over the graph for the data objects. We demonstrated the effectiveness of the proposed approach by utilizing spectral clustering and evaluating it over 20 Newsgroup and TREC datasets. The proposed approach is superior to the previous algorithms, especially for the correctness of cluster assignment. We plan to pursue this line of research to overcome the problem related with the instability of KL divergence in the text clustering problem.

Acknowledgements We express sincere gratitude to the reviewers for their careful reading of the manuscript and for providing valuable suggestions to improve the paper. This work is partially supported by the grant-in-aid for scientific research (No. 20500123) funded by MEXT, Japan.

References

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of 1998 ACM-SIGMOD* (pp. 94–105).

Akaike, H. (1973). Information theory and an extention of the maximum likelihood principle. In B. N. Petrov, & F. E. Csaki (Eds.), *2nd international symposium on information theory* (pp. 267–281).

Bekkerman, R., Sahami, M., & Learned-Miller, E. (2006). Combinatorial Markov random fields. In *Proceedings of the 17th European conference on machine learning (ECML-06)* (pp. 30–41).

Belkin, M., & Niyogi, P. (2002). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation, 15*, 1373–1396.

Chung, F. (1997). *Spectral graph theory*. American Mathematical Society.

Cover, T., & Thomas, J. (2006). *Elements of information theory*. Wiley.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(2), 1–38.

Dhillon, J., Mallela, S., & Modha, D. (2003). Information-theoretic co-clustering. In *KDD 2003* (pp. 89–98).
Dhillon, J., & Modha, D. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning, 42*, 143–175.
Diestel, R. (2006). *Graph theory*. Springer.
Elghazel, H., Kheddouci, H., Deslandres, V., & Dussauchoy, A. (2008). A graph b-coloring framework for data clustering. *Journal of Mathematical Modelling and Algorithms, 7*(4), 389–423.
Elghazel, H., Yoshida, T., Deslandres, V., Hacid, M., & Dussauchoy, A. (2007). A new greedy algorithm for improving b-coloring clustering. In *Proc. of the 6th workshop on graph-based representations* (pp. 228–239).
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of KDD-96* (pp. 226–231).
Frey, B. J. (1998). *Graphical models for machine learning and digital communication*. MIT Press.
Ghosh, J. (2003). *Scalable clustering* (pp. 341–364). Lawrence Erlbaum Associates.
Guénoche, A., Hansen, P., & Jaumard, B. (1991). Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *Journal of Classification, 8*, 5–30.
Guha, S., Rastogi, R., & Shim, K. (1998). Cure: An efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD conference* (pp. 73–84).
Hacid, H., & Yoshida, T. (2010). Neighborhood graphs for indexing and retrieving multidimensional data. *Journal of Intelligent Information Systems, 34*, 93–11.
Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Clustering validity checking methods: Part II. *ACM SIGMOD Record, 31*(3), 19–27.
Hansen, P., & Delattre, M. (1978). Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association, 73*, 397–403.
Hartigan, J., & Wong, M. (1979). Algorithm AS136: A k-means clustering algorithm. *Journal of Applied Statistics, 28*, 100–108.
Hartuv, E., & Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters, 76*, 175–181.
Irving, W., & Manlov, D. F. (1999). The b-chromatic number of a graph. *Discrete Applied Mathematics, 91*, 127–141.
Jain, A., Murty, M., & Flynn, T. (1999). Data clustering: A review. *ACM Computing Surveys, 31*, 264–323.
Li, T., Ma, S., & Ogihara, M. (2004). Entropy-based criterion in categorical clustering. In *Proceedings of the 21st ICML (ICML-04)* (pp. 536–543).
Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions On Pattern Analysis and Machine Intelligence, 24*(12), 1650–1654.
Muhlenbach, F., & Lallich, S. (2009). A new clustering algorithm based on regions of influence with self-detection of the best number of clusters. In *Proc. of 2009 IEEE international conference on data mining (ICDM'09)* (pp. 884–889).
Ng, R., & Han, J. (2002). Clarans: a method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering, 14*(5), 1003–1016.
Ogino, H., & Yoshida, T. (2010). Toward improving re-coloring based clustering with graph b-coloring. In *Proceedings of PRICAI-2010* (pp. 206–218).
Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proc. of the 30th annual meeting of the Association for Computational Linguistics* (pp. 183–190).
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.
Quinlan, J. R. (1993). *C4.5: Programs For machine learning*. Morgan Kaufmann.
Rissanen, J. (1978). Modeling by shortest data description methods in instance-based learning and data mining. *Automatica, 14*, 465–471.
Ristad, E. (1995). *A natural law of succession*. Technical Report CS-TR-495-95, Princeton University.
Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*(22), 2323–2326.
Slonim, N. (2002). *The information bottleneck: Theory and applications*. PhD thesis, Hebrew University.
Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *SIGIR-02* (pp. 129–136).
Slonim, N., & Tishby, N. (2000). Agglomerative information bottleneck. In *Advances in neural information processing systems (NIPS)* (Vol.12, pp. 617–623).
Stoer, M., & Wagner, F. (1997). A simple min-cut algorithm. *Journal of ACM, 44*(4), 585–591.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research, 3*(3), 583–617.
Tenenbaum, J., de Silva, J., & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(22), 2319–2323.
Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In *Proc. of the 37th allerton conference on communication and computation* (pp. 368–377).
Toussaint, G. T. (2005). Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. *International Journal of Computational Geometry Applications, 15*(2), 101–150.
Urquhart, R. (1982). Graph theoretical clustering based on limited neighbourhood sets. *Pattern Recognition, 15*(3), 173–187.
von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395–416.
Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers, 20*, 68–86.