Information Sciences 316 (2015) 87–106

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

CDIM: Document Clustering by Discrimination Information Maximization

Malik Tahir Hassan [a], Asim Karim [b], Jeong-Bae Kim [c], Moongu Jeon [a,*]

[a] School of Information and Communications, Gwangju Institute of Science and Technology, South Korea
[b] Department of Computer Science, SBASSE, Lahore University of Management Sciences, Pakistan
[c] Department of System Management, Pukyong National University, South Korea

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form 5 March 2015
Accepted 9 April 2015
Available online 15 April 2015

Keywords:
Document clustering
Discrimination information
Semantic relatedness
Relative risk
Cluster understanding

ABSTRACT

Ideally, document clustering methods should produce clusters that are semantically relevant and readily understandable as collections of documents belonging to particular contexts or topics. However, existing popular document clustering methods often ignore term-document corpus-based semantics while relying upon generic measures of similarity. In this paper, we present CDIM, an algorithmic framework for partitional clustering of documents that maximizes the sum of the discrimination information provided by documents. CDIM exploits the semantic that term discrimination information provides better understanding of contextual topics than term-to-term relatedness to yield clusters that are describable by their highly discriminating terms. We evaluate the proposed clustering algorithm using well-known discrimination/semantic measures including Relative Risk (RR), Measurement of Discrimination Information (MDI), Domain Relevance (DR), and Domain Consensus (DC) on twelve data sets to prove that CDIM produces high-quality clusters comparable to the best methods. We also illustrate the understandability and efficiency of CDIM, suggesting its suitability for practical document clustering.

1. Introduction

Data clustering is one of the most widely used task in data mining due to its capability for summarizing large data collections. The objective of data clustering methods is to find groups of data objects that are related to one another within a group and are unrelated to objects in other groups. These methods, which are unsupervised in nature, often optimize an objective function that captures an appropriate notion of clustering (e.g. maximize similarity between objects within groups, and dissimilarity of objects between groups).

Textual document clustering discovers groups of related documents in large document collections. Its importance has grown significantly over the years as the world moves toward a paperless environment and the Web continues to dominate our lives. Efficient and effective document clustering methods can help us with better document organization (e.g. digital libraries, corporate documents) as well as quicker and improved information retrieval (e.g. online search).

Besides the need for efficiency, document clustering methods should be able to handle the large term space of document collections to produce semantically relevant and readily understandable clusters. These requirements are often not satisfied in popular clustering methods. For example, in *K*-means clustering [29], documents are compared in the term space, which is

* Corresponding author. Tel.: +82 62 715 2406.
E-mail address: mgjeon@gist.ac.kr (M. Jeon).

http://dx.doi.org/10.1016/j.ins.2015.04.009
0020-0255/© 2015 Elsevier Inc. All rights reserved.

typically sparse, using generic similarity measures without considering the term-document semantics other than their vectorial representation in space. Moreover, it is not straightforward to interpret and understand the clusters formed by K-means clustering; the similarity of a document to its cluster's mean provides little understanding of the document's context or topic.

In this paper, we present a document clustering framework based on discrimination information maximization (CDIM). The CDIM algorithm was introduced in our previous work [25]. This paper extends the preliminary work by presenting CDIM as a homogeneous document clustering framework enabling the use of different discrimination and relevance measures. Other major contributions include the psycholinguistics based motivation to use discrimination information for document clustering, presentation of CDIM variants, proof of convergence, comprehensive clustering quality and understanding evaluation on more data sets and against more competitors.

The iterative procedure of CDIM repeatedly projects documents onto a K-dimensional discrimination information space and assigns documents to the cluster along whose axis they have the largest value. The discrimination information space is defined by term discrimination information estimated from the labeled document collection produced in the previous iteration. This procedure maximizes the sum of discrimination information provided by all documents. A key advantage of using term discrimination information is that each cluster can be identified by a list of highly discriminating terms. These terms can also be thought of as units of thought describing a cluster in the document collection. As a result of this semantic interpretation, the clusters produced by CDIM are understandable by their discriminating terms. Since CDIM is posed as an optimization problem, there is room to apply different measures in objective function. We present results using Relative Risk (RR) [36], Measurement of Discrimination Information (MDI) [8], Domain Relevance (DR) and Domain Consensus (DC) [43]. Other variations of CDIM that we implement include CDIM using repeated bisection. We evaluate the performance of CDIM on twelve popular text data sets. In clustering quality evaluation, CDIM is found to produce high quality clusters that are significantly better than those produced by flat or partitional methods like spectral clustering [12], non-negative matrix factorization (NMF) [52], K-means and its variants [59]. Performance of CDIM is also significantly better than the hierarchical methods HFTC [6] and Rank-2 NMF [34], and is comparable to the famous hierarchical methods FIHC [18] and UPGMA [30]. We demonstrate that CDIM provides better understanding of clusters than FIHC and UPGMA. The quality of clustering is determined using BCubed F-measure [1] which combines BCubed precision and BCubed recall. F-measure is also calculated in order to compare our results with existing published results. Our results suggest the practical suitability of CDIM for clustering and understanding of document collections.

The rest of the paper is organized as follows. We discuss the related works and the motivation for our method in Section 2. Our document clustering method, CDIM, is described in detail in Section 3. Variants of CDIM are presented in Section 4. Section 5 presents our experimental setup. Section 6 discusses the results of our experiments. We conclude and give some future directions in Section 7.

2. Motivation and related work

In this section, we describe the motivation and discuss the related works to our discrimination information based document clustering framework. For convenience, we divide this section into two subsections. The first subSection 2.1 discusses use of discrimination information in data processing and its significance in document analysis, and the second subSection 2.2 discusses related clustering methods.

2.1. Discrimination information

Discrimination, or association as its opposite concept, is a fundamental concept in information processing [46]. It is central to many data mining tasks such as classification and feature selection. Measures of discrimination information come from statistics and information theory. Common measures include relative risk, odds ratio, risk difference, information gain, and Kullback–Leibler divergence. These measures are corpus-based, i.e., they are estimated from a data collection.

In recent years, there has been growing interest in using statistically sound measures in data mining [36,37]. In the biomedical domain, on the other hand, measures like relative risk and odds ratio have been used for a long time for cohort studies and factor analysis [26,35]. In text processing, such measures have been used primarily for feature selection [13]. More recently, measures like relative risk and information gain have been used to quantify the discrimination information provided by terms for text classification purposes [31,40]. These works highlight the suitability of building learning models from term discrimination information.

The semantics of term discrimination information has been discussed by Cai [10]. They present a theoretical framework to estimate semantic relatedness between terms and how this relatedness can help in identifying a term's strongest support category among all categories in a document collection. In a similar context, Xu et al. [53] measure SDC (semantic discrimination capability) of association relations between terms, and illustrate its applicability to document clustering.

In the psycholinguistics domain, it has been shown that humans are more likely to associate terms with their respective contexts or topics rather than associate terms with other terms in a given context [22,41]. Thus, term-to-term semantic relations (e.g. synonymy), which are more commonly used in text analysis, provide only indirect information about the contexts in which terms are used. Furthermore, this suggests that term discrimination information can be used to identify groups of related documents. The connection between document clustering and term discrimination is illustrated in Fig. 1. This figure shows three clusters together with the discriminating terms in each cluster. Documents with similar usage of discriminating terms fall in one cluster, and these documents typically represent a coherent topical context.

2.2. Clustering

Data clustering has been studied extensively for over 50 years with widespread usage in applications across different disciplines [28,29]. This popularity stems from the intuitive objective of clustering; finding groups of related objects. Even so, developments in data clustering continue specially in applications areas like document clustering. In general, data clustering can be partitional or hierarchical in nature [49]. Partitional clustering finds K clusters (where K is an input parameter) by optimizing an objective function that captures an appropriate notion of clustering, while (agglomerative) hierarchical clustering creates a tree of clusters by performing local merge operations. Since our method is partitional in nature, we will focus primarily on this type of clustering methods.

Content-based document clustering has also been studied since the 1960s [45,3]. Document clustering continues to be a challenging problem [48] because of (1) the high dimensionality of the term-document space, (2) the sparsity of the documents in the term-document space, and (3) the difficulty of incorporating appropriate term-document semantics for improved clustering quality and understandability. Moreover, real-world document clustering often involves large document collections thus requiring the clustering method to be efficient.

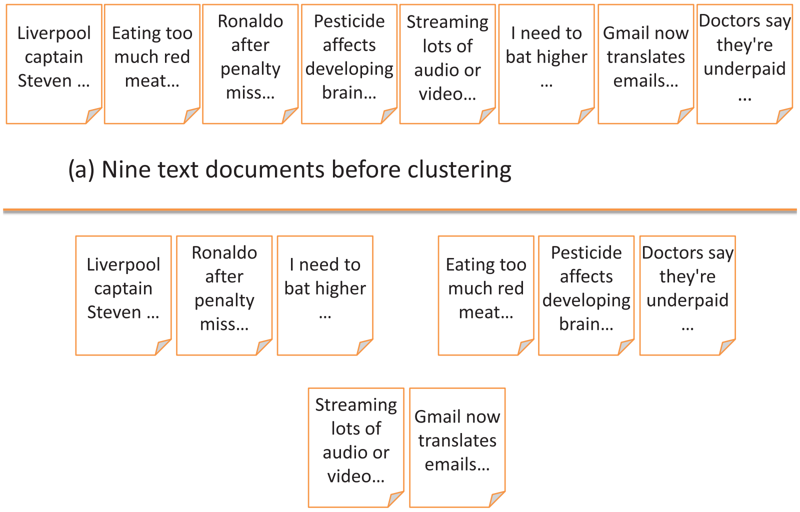
The K-means algorithm continues to be popular for document clustering due to its efficiency and ease of implementation [49]. It is a partitional clustering method that optimizes an objective function via iterative two-step procedure. Usually, documents are represented by terms that are weighted by term-frequency or term-frequency-inverse-document-frequency values, and documents are compared in the term space by the cosine similarity measure. Several clustering objective functions can be optimized [58]. These functions can be based on notions of cohesion, separation, or both cohesion and separation (hybrid) of clusters. It has been reported that the objective function of the traditional K-means algorithm which maximizes the similarity of documents to their cluster mean produces reliable clusterings [58]. The Repeated Bisection clustering method, which splits clusters into two until the desired number of clusters are obtained, has been shown to produce better clusterings especially when K is large (greater than 20) [47]. These K-means based methods are efficient and accurate for many practical applications. Their primary shortcoming is poor interpretability of the clusters where the cluster mean vector is often not a reliable indicator of the documents in a cluster. A topic based clustering approach has been presented in [39]. In this approach, initial cluster centers or seeds are selected and K-means++ algorithm [4] is applied. CDIM experiments are conducted using random seeding in our work. Careful seeding will further improve the performance of CDIM potentially [4].

Documents can be grouped and visualized on different levels of abstractions. Thus many interesting hierarchical methods have been applied to cluster documents [59]. Agglomerative clustering using Un-weighted Pair Group Method with Arithmetic Mean (UPGMA) [30], Frequent term-based text clustering (HFTC) [6] and Frequent Itemset-based Hierarchical clustering (FIHC) [18] are some of the most prominent works in this category. There are hierarchical document clustering algorithms that handle dynamic and overlapping data [21] in literature. These are considered out of scope for this work as CDIM is a flat, static and hard-clustering algorithm by nature.

Frequent pairs and sequence based approaches for document clustering also give good results [2,38,56]. A sequence is different from an itemset as the order of terms is preserved in a sequence and not in an itemset. These n-gram, itemset and sequence based approaches are generally more computationally expensive. Our CDIM approach gives very competitive results by using just unigrams. Extension of CDIM to use item sets is a straight forward and promising direction.

Some researchers have used external knowledge bases to semantically enrich the document representation for document clustering [38,27,57,42]. In [27], Wikipedia's concepts and categories are adopted to enhance the document representation, while in [38,57,42] several ontology-based (e.g. WordNet) term relatedness measures are evaluated for semantically smoothing the document representation. In all of these works, it has been shown that the quality of clusterings produced improves over the baseline ("bag of words") document representation. However, extracting information from knowledge bases is computationally expensive. Furthermore, most of these approaches suffer from the same shortcomings of K-means regarding cluster understandability.

The challenge of high dimensional data clustering, including that of document clustering, has received much interest in recent years [33,12]. These methods try to find clusters in lower dimensional subspace(s) of the original term space or a transformed space. One way to find clusters in transformed spaces is through Non-Negative Matrix Factorization (NMF). NMF approximates the term-document matrix by the product of term-cluster and document-cluster matrices [52]. Extensions to this idea, with the goal of improving the interpretability of the extracted clusters, have also been proposed [34,51,9]. Spectral clustering methods [12] exploit pairwise similarities of documents instead of similarity to centers like in K-means. These methods compute K eigenvectors of a Laplacian matrix to generate clusters. Document clustering in a low dimension space has also been presented by [55]. The claim is that correlation as the similarity measure is more suitable than Euclidean distance to detect the intrinsic geometrical structure of the document space. Nonetheless, these methods are restricted by their focus on approximation rather than semantically useful clusters. Our method, on the other hand, focusses on finding highly discriminating clusters where documents having similar behaviors (discrimination score for a cluster) are grouped together.


Liverpool captain Steven ...
Eating too much red meat...
Ronaldo after penalty miss...
Pesticide affects developing brain...
Streaming lots of audio or video...
I need to bat higher ...
Gmail now translates emails...
Doctors say they're underpaid ...
(a) Nine text documents before clustering
Liverpool captain Steven ...
Ronaldo after penalty miss...
I need to bat higher ...
Eating too much red meat...
Pesticide affects developing brain...
Doctors say they're underpaid ...
Streaming lots of audio or video...
Gmail now translates emails...

(b) Nine text documents after clustering into three groups

Fig. 1. Illustrating document clustering where unlabeled documents are partitioned into three groups based on discriminating terms.

Another way to find clusters in transformed spaces is to combine clustering methods with dimensionality reduction techniques. Unsupervised dimensionality reduction techniques have been proposed for document clustering, and a comparison of four such techniques is given in [50]. It is reported that when Latent Semantic Indexing (LSI) is combined with *K*-means algorithm, reliable clusterings are obtained. The combination of *K*-means with Linear Discriminant Analysis (LDA), which is a supervised dimensionality reduction technique, has also been investigated [14,17]. These methods iteratively perform clustering in low-dimensional spaces found by LDA. Our method uses a similar iterative procedure of finding clusters in a low-dimensional space. However, its procedure is more efficient than the methods that use LDA and LSI. In addition, dimensionality reduction is also provided automatically as the lesser discriminating terms can be ignored using the term selection parameter.

A related concept is that of co-clustering [16] which clusters documents and terms simultaneously at all stages. It produces clusters of documents that show similar behavior across clusters of terms, or vice versa. CDIM in comparison is a one-way clustering method that produces clusters of documents, but it also returns a ranking of terms in every cluster. Moreover, this ranking of terms is done automatically based on the discrimination information in the document collection. As such, CDIM's results are more readily understandable than those produced by co-clustering methods.

3. CDIM – Clustering by Discrimination Information Maximization

CDIM (Clustering by Discrimination Information Maximization) is an iterative partitional document clustering framework that finds *K* groups of documents in a *K*-dimensional discrimination information space transformed from the *M*-dimensional input space. It does this by following an efficient two-step procedure of document projection and assignment with the goal of maximizing the sum of documents' discrimination scores. CDIM's clusters are describable by highly discriminating terms related to the context/topic of the documents in the cluster. We start our presentation of CDIM by formally stating the problem.

3.1. Problem statement

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \Re^{M\times N}$ be the term-document matrix in which $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{Mi}]^T$ is the *i*th document represented by an *M*-dimensional vector (*i*th column of matrix $\mathbf{X}$). *M* is the total number of distinct terms in the *N* documents. The weight of term x_j in document $\mathbf{x}_i$, denoted by x_{ji}, is equal to the count of the term x_j in the document $\mathbf{x}_i$.

Our goal is to find *K* (usually in practice $K \ll \min\{M, N\}$) clusters C_k $(k = 1, 2, \ldots, K)$ of documents such that if a document $\mathbf{x} \in C_k$ then $\mathbf{x} \notin C_j, \forall j \neq k$. Thus, we assume hard partitioning of the documents among the clusters; however, this assumption can be relaxed trivially in CDIM but we do not discuss this further in our current work. In addition to the cluster composition, we will also like to find significant describing terms for each cluster. Let T_k be the index set of significant terms for cluster *k*.

3.2. Clustering objective function

CDIM finds *K* clusters in the document collection by maximizing the sum of discrimination scores of documents for their respective clusters. If we denote the discrimination information provided by document $\mathbf{x}_i$ for cluster *k* by d_{ik} and the discrimination information provided by document $\mathbf{x}_i$ for all clusters but cluster k by $\bar{d}_{ik}$, then the discrimination score of document $\mathbf{x}_i$ for cluster k is defined as $\hat{d}_{ik} = d_{ik} - \bar{d}_{ik}$. CDIM's objective function can then be written as

$$J = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \delta_{ik}(d_{ik} - \bar{d}_{ik}) \tag{1}$$

where $\delta_{ik} = 1$ if document $\mathbf{x}_i$ is assigned to cluster k and zero otherwise. Document discrimination information (d_{ik} and $\bar{d}_{ik}$) is computed from term discrimination information that in turn is estimated from the current labeled document collection. These computations are discussed in the subsequent subsections.

Intuitively, CDIM seeks a clustering in which the discrimination information provided by documents for their cluster is higher than the discrimination information provided by them for the remaining clusters. It is not sufficient to maximize just the discrimination information of documents for their respective clusters as they may also provide high discrimination information for the remaining clusters.

The objective function J is maximized by using a greedy two-step procedure. In one step, given a cluster assignment defined by $\delta_{ik}, \forall i, k, J$ is maximized by estimating $d_{ik}, \forall i, k$ and $\bar{d}_{ik}, \forall i, k$ from the labeled document collection. This estimation is done using maximum likelihood estimation, described in Section 3.4. In the other step, given estimated discrimination scores $\hat{d}_{ik}, \forall i, k$ of documents, J is maximized by assigning each document to the cluster k for which the document's discrimination score is maximum. This two-step procedure continues until the change in J from one iteration to the next drops below a specified threshold value (convergence is formally proved in Section 3.9).

3.3. Term discrimination information

The discrimination information provided by a document is computed from the discrimination information provided by the terms in the document. The discrimination information provided by a term is quantified with a measure of discrimination information that is estimated from the labeled document collection. The following subsections show our formulation for two discrimination measures: Relative Risk (RR) [36] and Measurement of Discrimination Information (MDI) [8].

3.3.1. Relative Risk (RR)

Our first measure of term discrimination information is the Relative Risk (RR), which is popularly used in biomedical studies [36]. Relative risk of a term for cluster k over the remaining clusters is used as its discrimination information for cluster k. Mathematically, the discrimination information of term x_j for cluster k and term x_j for all clusters but k is given by

$$w_{jk} = \begin{cases} \frac{p(x_j|C_k)}{p(x_j|\bar{C}_k)} & \text{when } p(x_j|C_k) - p(x_j|\bar{C}_k) > t \\ 0 & \text{otherwise} \end{cases} \text{and} \tag{2}$$

$$\bar{w}_{jk} = \begin{cases} \frac{p(x_j|\bar{C}_k)}{p(x_j|C_k)} & \text{when } p(x_j|\bar{C}_k) - p(x_j|C_k) > t \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $p(x_j|C_k)$ is the conditional probability of term x_j in cluster k, $\bar{C}_k$ denotes all clusters but cluster k, and t is a term selection parameter that controls the exclusion of terms that provide insignificant discrimination information. A discussion on t is given in Section 3.5. The term discrimination information is either zero (no discrimination information) or greater than one with a larger value signifying higher discriminative power.

3.3.2. Measurement of Discrimination Information (MDI)

Our second measure of term discrimination information, called measurement of discrimination information (MDI), is adopted from Cai and Rijsbergen [8]. This work studies semantic relatedness between terms using term discrimination information and proposes three term discrimination information measures to quantify and characterize the relatedness. The first and second measures, identified as $\mathbf{ifd}_{I1\Sigma}$ and $\mathbf{ifd}_{I2\Sigma}$, quantify discrimination as a distribution divergence between category 1 and combined data (identified in the subscript with 1Σ), and category 2 and combined data (identified in the subscript with 2Σ), respectively. The third measure, identified as $\mathbf{ifd}_K$, is a linear combination of the first two measures. In our context of document clustering, categories 1 and 2 correspond to clusters C_k and $\bar{C}_k$, respectively. With this interpretation, these measures are defined as [8]:

$$\mathbf{ifd}_{I1\Sigma}(x_j) = p(x_j|C_k) log \frac{p(x_j|C_k)}{\lambda_1 p(x_j|C_k) + \lambda_2 p(x_j|\overline{C}_k)} \tag{4}$$

$$\mathbf{ifd}_{I2\Sigma}(x_j) = p(x_j|\bar{C}_k) log \frac{p(x_j|\bar{C}_k)}{\lambda_1 p(x_j|C_k) + \lambda_2 p(x_j|\bar{C}_k)} \tag{5}$$

$$\mathbf{ifd}_K = \lambda_1 \mathbf{ifd}_{I1\Sigma} + \lambda_2 \mathbf{ifd}_{I2\Sigma} \tag{6}$$

λ_1 and λ_2 are the prior probabilities of C_k and $\bar{C}_k$, respectively. The discrimination of a term x_j is characterized via the following inequalities [8]:

$$\psi_1 = \lambda_1 \mathbf{ifd}_{l1\Sigma} - \lambda_2 |\mathbf{ifd}_{l2\Sigma}| > 0 \tag{7}$$

$$\psi_2 = \lambda_2 \mathbf{ifd}_{l2\Sigma} - \lambda_1 |\mathbf{ifd}_{l1\Sigma}| > 0 \tag{8}$$

If the first inequality is satisfied, the term supports category 1 more than category 2 (or, in our context, cluster C_k more than the rest) and when the second inequality is satisfied, the term supports category 2 more than category 1.

Using inequalities 7 and 8, the discrimination term weights for CDIM, w_{jk} and $\bar{w}_{jk}$ are given as

$$w_{jk} = \begin{cases} \psi_1 & \text{when } \psi_1 > t \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$\bar{w}_{jk} = \begin{cases} \psi_2 & \text{when } \psi_2 > t \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

3.4. Estimation and smoothing

The various cluster-conditional and cluster prior probabilities used in the definition of term discrimination information presented in the preceding section are estimated from the currently labeled document collection using maximum likelihood estimation (MLE). MLE is a straightforward procedure but is prone to over-fitting and can produce zero probability estimates. As such, appropriate smoothing of the estimates is required for robust processing. This is especially important for iterative clustering methods since cluster sizes can vary significantly from iteration to iteration.

There exist many smoothing techniques with different levels of sophistication [11]. The more sophisticated techniques, like Jelinek-Mercer, Katz, Witten–Bell, and Kneser–Ney, combine higher order n-gram models with lower order n-gram models for improved probability estimation. The most popular simpler techniques include additive smoothing and Good-Turing smoothing. Additive smoothing has been criticized for its poorer performance in text analytics [19,44,24]. Since CDIM is based on a 1-gram model we adopt the Good-Turing estimator [20].

According to this technique, the non-zero probabilities are reduced by a factor of $1 - E(1)/T$ and zero probabilities get equal share from $E(1)/T$. Here $E(1)$ is the expected number of terms that occur once and T is the corpus size.

3.5. Relatedness of terms to clusters

The term selection parameter $t \geqslant 0$ in Eqs. (2), (3), (9) and (10) enables to remove the lesser related terms. As the value of t is increased from zero, fewer terms will have a high discrimination information. The index set of terms that provide significant discrimination information for cluster k (T_k) is defined as

$$T_k = \{j | w_{jk} > 0, \forall j\}. \tag{11}$$

These terms and their discrimination information provide a good understanding of the context of documents in cluster k in contrast with those in other clusters in the document collection. In general, $T_k \cap T_j \neq \emptyset, \forall j \neq k$. That is, there may be terms that provide significant discrimination information for more than one cluster. Also, depending on the value of t, there may be terms that do not provide significant discrimination information for all clusters.

In a study discussed in [41], it has been shown that humans comprehend text by associating terms with particular contexts or topics. These relationships are different from the traditional lexical relationships (e.g. synonymy, antonymy, etc.), but are more fundamental in conveying meaning and understanding. Recently, it has been shown that the degree of relatedness of a term to a context is proportional to the term's discrimination information for that context in a corpus [8]. Given these studies, we can consider all terms in T_k to be related to cluster k and the strength of this relatedness is given by the term's discrimination information. This is an important characteristic of CDIM whereby each cluster's context is describable by a set of related terms. Furthermore, these terms and their weights (discrimination information) define a K-dimensional space in which documents are comparable by their discrimination information.

3.6. Document discrimination information

A document $\mathbf{x}_i$ is describable by the terms it contains. Each term x_j in the document vouches for the context or cluster k according to the value of the term's discrimination information w_{jk}. Equivalently, each term x_j in the document has a certain degree of relatedness to context or cluster k according to the value w_{jk}. The discrimination information provided by document $\mathbf{x}_i$ for cluster k can be computed as the average term discrimination information for cluster k:

$$d_{ik} = \frac{\sum_{j \in T_k} x_{ji} w_{jk}}{\sum_j x_{ji}}. \tag{12}$$

A similar expression can be used to define $\bar{d}_{ik}$. The document discrimination information d_{ik} can be thought of as the relatedness (discrimination) of document $\mathbf{x}_i$ to cluster k. The document discrimination score is given by $\hat{d}_{ik} = d_{ik} - \bar{d}_{ik}$; the larger this value is, the more likely that document $\mathbf{x}_i$ belongs to cluster k. Note that a term contributes to the discrimination information of document $\mathbf{x}_i$ for cluster k only if it belongs to T_k and it occurs in document $\mathbf{x}_i$. If such a term occurs multiple times in the document then each of its occurrence contributes to the discrimination information. Thus, the discrimination information of a document for a particular cluster increases with the increase in occurrences of highly discriminating terms for that cluster.

3.7. Algorithm

CDIM is an EM variation of the K-means with an objective function that maximizes the discrimination information of documents for clusters. The algorithm can be presented more compactly in matrix notation. CDIM's algorithm, which is outlined in Algorithm 1, is described next.

Let $\mathbf{W}$ ($\bar{\mathbf{W}}$) be the $M \times K$ matrix formed from the elements w_{jk} $(\bar{w}_{jk}), \forall j, k$, $\hat{\mathbf{D}}$ be the $N \times K$ matrix formed from the elements $\hat{d}_{ik}, \forall i, k$, and $\mathbf{R}$ be the $N \times K$ matrix formed from the elements $\delta_{ik}, \forall i, k$. At the start, each document is assigned to one of the K randomly selected seeds using cosine similarity, thus defining the matrix $\mathbf{R}$. Then, a loop is executed consisting of two steps. In the first step, the term discrimination information matrices ($\mathbf{W}$ and $\bar{\mathbf{W}}$) are estimated from the term-document matrix $\mathbf{X}$ and the current document assignment matrix $\mathbf{R}$. The second step projects the documents onto the relatedness or discrimination score space to create the discrimination score matrix $\hat{\mathbf{D}}$. Mathematically, this transformation is given by

$$\hat{\mathbf{D}} = (\mathbf{X}\Sigma)^T(\mathbf{W} - \bar{\mathbf{W}}) \tag{13}$$

where Σ is a $N \times N$ diagonal matrix defined by elements $\sigma_{ii} = 1/\sum_j x_{ji}$. The matrix $\hat{\mathbf{D}}$ represents the documents in the K-dimensional discrimination score space.

Documents are re-assigned to clusters based on their discrimination scores. A document $\mathbf{x}_i$ is assigned to cluster k if $\hat{d}_{ik} > \hat{d}_{ij}, \forall j \neq k$ (ties are broken arbitrarily). In matrix notation, we write this operation as

$$\mathbf{R} = \text{maxrow}(\hat{\mathbf{D}}) \tag{14}$$

where 'maxrow' is an operator that works on each row of $\hat{\mathbf{D}}$ and returns a 1 for the maximum value and a zero for all other values. The processing of Eqs. (13) and (14) are repeated until the absolute difference in the objective function becomes less than a specified small value. The objective function J is computed by summing the maximum values from each row of matrix $\hat{\mathbf{D}}$.

The algorithm outputs the final document assignment matrix $\mathbf{R}$ and the final term discrimination information matrix $\mathbf{W}$.

Algorithm 1. CDIM – Document Clustering by Discrimination Information Maximization.

```
Require: $\mathbf{X}$ (term-document matrix), $K$ (No. of clusters)
$\mathbf{R}^{(0)} \leftarrow$ initial assignment of documents to clusters
$\tau \leftarrow 0$
$J^{(0)} \leftarrow 0$
repeat
    $\mathbf{W}^{(\tau)}, \bar{\mathbf{W}}^{(\tau)} \leftarrow$ term discrimination info estimated from $\mathbf{X}$ and $\mathbf{R}^{(\tau)}$ (Eqs. (2) and (3) or Eqs. (9) and (10))
    $\hat{\mathbf{D}}^{(\tau+1)} \leftarrow (\mathbf{X}\Sigma)^T(\mathbf{W}^{(\tau)} - \bar{\mathbf{W}}^{(\tau)})$
    $\mathbf{R}^{(\tau+1)} \leftarrow \text{maxrow}(\hat{\mathbf{D}}^{(\tau+1)})$
    $J^{(\tau+1)} \leftarrow$ sum of max discrimination scores from each row of $\hat{\mathbf{D}}^{(\tau+1)}$
    $\tau \leftarrow \tau + 1$
until $(|J^{(\tau)} - J^{(\tau-1)}| < \epsilon)$
return $\mathbf{R}$ (document assignment matrix), $\mathbf{W}$ (term discrimination info matrix)
```

3.8. Key characteristics of CDIM

We highlight some characteristics and properties of our document clustering method below.

1. The computational time complexity of CDIM is $O(KMNI)$ where I is the number of iterations required to reach the final clustering. Thus, the computational time of CDIM depends linearly on the clustering parameters.

2. CDIM does not require the specification of a document-to-document similarity or dissimilarity measure. Documents are projected onto a K dimensional discrimination score space in such a way that the relevance of a document to a cluster is given by its value on the corresponding axis – the larger this value is, the more relevant the document is to that cluster.
3. In addition to outputting the document assignment matrix, CDIM also outputs the term discrimination information matrix (**W**). This matrix identifies the significant discriminating terms for each cluster and quantifies their discrimination information or relatedness for/to each cluster. This information is valuable for understanding the context of documents in each cluster.
4. CDIM includes a natural way for term selection via the term selection parameter t (see Eqs. (2) and (3)). By increasing the value of t, terms providing little discrimination information (potential noise) can be removed easily to further improve clustering performance.
5. The two-step procedure of document assignment based on discrimination information estimation from the labeled documents produced by the previous iteration represents a naive semi-supervised learning approach [54]. In [54], this approach is shown to converge smoothly in about 20 iterations. Formal proof of convergence is presented in the next section.

We illustrate and discuss some of these characteristics in our experiments i.e. Section 5.

3.9. Convergence of CDIM

CDIM is a two-step iterative algorithm like K-means and Expectation Maximization (EM). The two steps of CDIM are calculation of discrimination scores and assignment of documents to clusters. In K-means, a cluster is represented by its mean or centroid. Whereas in CDIM a cluster is represented by its significant discriminating terms. The iterations in K-means try to stabilize the cluster means; likewise, the iterations in CDIM try to stabilize the set of significant discriminating terms. Below, we prove that CDIM converges/terminates by proving that both of its steps monotonically increase the objective function, and the value of objective function is bounded from above.

Theorem 3.1. *CDIM algorithm converges to local maxima.*

Proof. Consider CDIM's objective function (Eq. (1)). It is the sum of discrimination scores of documents for their respective clusters, where the discrimination score of document $\mathbf{x}_i$ for cluster k is $\hat{d}_{ik} = d_{ik} - \bar{d}_{ik}$. The objective function can be written as the sum of K clusters' documents' discrimination scores:

$$J = \sum_{\mathbf{x}_i \in C_1} \delta_{i1}(\hat{d}_{i1}) + \sum_{\mathbf{x}_i \in C_2} \delta_{i2}(\hat{d}_{i2}) + \cdots + \sum_{\mathbf{x}_i \in C_K} \delta_{iK}(\hat{d}_{iK}) \tag{15}$$

Let $C_1, C_2, \ldots, C_k$ be our current clustering with objective function value J_t. Based on current labels we calculate discrimination scores and find that a document $\mathbf{x}_n \in C_i$ needs relabeling, i.e.,

$$\hat{d}_{nj} > \hat{d}_{ni}, \forall i \neq j \tag{16}$$

This implies that document n has higher discrimination score for cluster j than all other clusters (including the current cluster i). Guided by our new discrimination score calculations, we assign the document n to cluster j. This assignment results in a decrement of $\hat{d}_{ni}$ in discrimination score of cluster i but an increment of $\hat{d}_{nj}$ in discrimination score of cluster j.

Summing the discrimination scores of all clusters now produces new objective function value J_{t+1} and

$$J_{t+1} > J_t \tag{17}$$

because of condition 16.

The migration of document n from cluster i to cluster j produces three possible types of terms:

(1) Terms of document n whose weights increase in cluster j; let this subset of terms be S_j.
(2) Terms of document n whose weights decrease in cluster j; let this subset of terms be S_i.
(3) Terms in document n with unchanged weights.

But we see that

$$\sum_{a \in S_j} w_{aj} > \sum_{a \in S_i} w_{ai} \tag{18}$$

because of 17. Therefore, there can be terms with decreased weights but overall to increase J_{t+1} the total contribution of term weights is monotonically increased. And, for next iteration $J_{t+2} \geqslant J_{t+1}$.

Since J is bounded above by the sum of discrimination scores of best possible clustering, CDIM will converge, although it may converge to some local maxima. □

4. Variations of CDIM

CDIM is an algorithmic framework that can allow several specific variants. Different choice of term discrimination information in CDIM has already been discussed; here, we present these and other variations of CDIM that are evaluated in this work.

4.1. CDIM-RR and CDIM-MDI

CDIM-RR and CDIM-MDI are two versions of CDIM defined by using the discrimination measures Relative Risk (RR) and measurement of discrimination information (MDI) for calculation of term weights respectively, described in Section 3.3. Document discrimination information is then calculated using Eq. (12). The CDIM algorithm steps are then iterated to optimize objective function of Eq. (1).

CDIM-MDI-S is the symmetric version of CDIM-MDI which is obtained by assigning equal prior probabilities while calculating the measurement of discrimination information, as described in Section 3.3.2.

4.2. CDIM for repeated bisection

The original CDIM algorithm works in a k-way fashion, also called direct clustering. The data is initially partitioned into k clusters and then these partitioned are refined iteratively to reach an optimum value of the objective function. Another possible method is to split the data into two clusters and then keep on splitting a candidate cluster until the desired number of clusters i.e. k, is obtained. This type of algorithm is known as repeated bisection (RB) in literature [47].

In repeated bisection version of CDIM, we select the largest cluster at each step and split it into two clusters using CDIM algorithm. Selecting largest cluster as bisection candidate is simple and using other complex selection methods do not give significantly better results in general [47,59]. Repeated bisection results are presented for both discrimination measures i.e. relative risk and MDI, yielding the versions CDIM-RR-RB and CDIM-MDI-RB respectively.

4.3. CDIM using information retrieval measures (DR and DC)

Besides using discrimination measures in our clustering method, we experiment with relevance measures too. We combine Domain Relevance (DR) and Domain Consensus (DC) as proposed by [43], to assign cluster relevance weights to terms.

In our clustering context, domain relevance of a term x_j to a cluster C_k is calculated as

$$DR_{jk} = \frac{p(x_j|C_k)}{\max_{1\leqslant v\leqslant n} p(x_j|C_v)}. \tag{19}$$

Domain consensus is computed to see the distributed use of a term in a cluster. A term with wide usage in a cluster i.e. present in many documents, is more relevant to the cluster than a term that occurs a large number of times in small number of documents in the cluster. Domain consensus for a term x_j present in the documents belonging to a cluster C_k is computed as

$$DC_{jk} = \sum_{\mathbf{x}_i \in C_k} p(x_j|\mathbf{x}_i) log \frac{1}{p(x_j|\mathbf{x}_i)}. \tag{20}$$

The two measures are then combined to calculate the term weights

$$w_{jk} = \alpha DR_{jk} + \beta DC_{jk}^{norm}. \tag{21}$$

After some experimentation, the values for α and β are chosen to be 0.9 and 0.3 respectively as recommended by [43]. DC is normalized by the number of documents in the cluster.

These term weights w_{jk} are pooled to get document weights d_{ik}. The objective function for CDIM now maximizes the domain relevance and domain consensus as

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} \delta_{ik} d_{ik} \tag{22}$$

where $\delta_{ik} = 1$ if document $\mathbf{x}_i$ is assigned to cluster k and zero otherwise. We name this version of CDIM that uses domain relevance and domain consensus as CDIM-DRDC.

5. Experimental setup

We conduct extensive experimental evaluations of our document clustering method. Our evaluations comprise of three sets of experiments. First, we evaluate the clustering quality of CDIM and compare it with other clustering methods on twelve text data sets. Second, we illustrate the understanding that is provided by CDIM clustering. Third, we evaluate certain implementation issues including term selection, convergence, and probability estimation. The results of these experiments are given in the next section. Here, we describe our experimental setup.

5.1. Data sets

We conduct experiments on twelve standard text data sets of different sizes, contexts, and complexities. The key characteristics of these data sets are given in Table 1 and Fig. 2. Data sets 1 (stopword removal) and 3 to 12 (stopword removal and stemming) are available in preprocessed formats, while we perform stopword removal and stemming of data set 2. Data set 1 is obtained from the Internet Content Filtering Group's web site,[1] data set 2 is available from a Cornell University Web page,[2] and data sets 3 to 11 are obtained from Karypis Lab, University of Minnesota.[3] Brief description of each data set is as below.

The pu data set contain e-mails received by a particular user labeled as either spam or non-spam. The movie data set contain reviews of movies from the Internet Movie Database (IMDB). Each document is labeled as either a positive or a negative review. The hitech data set has newspaper articles belonging to one of six categories: computers, electronics, health, medical, research, and technology. The reviews data set contain articles about food, movies, music, radio, and restaurants. Both of these two data sets are derived from the San Jose Mercury newspaper articles that are distributed as part of the TREC (Text Retrieval Conference) collection (TIPSTER Vol. 3).

The seven and ten category data sets tr31 and tr41 are derived from TREC-6 and TREC-7 collections.[4] The categories in these data sets correspond to the queries and the documents that are judged most relevant to them.

The data sets re0 and re1 are taken from Reuters-21578 text categorization test collection distribution 1.0. The labels are divided into two sets producing two different data sets. The documents having single label only are used for both data sets. The data set reuters is also a subset of Reuters-21578, and is publicly available.[5]

The ohscal data set is derived from the OHSUMED collection of medical publications. It contain documents from ten categories: antibodies, carcinoma, DNA, in vitro, molecular sequence data, pregnancy, prognosis, receptors, risk factors, and tomography. The wap data set is obtained from the WebACE project and each document corresponds to a web page listed in the subject hierarchy of Yahoo!. Classic data set consists of four different document categories: CACM, CISI, CRAN, and MED.

5.2. Comparison methods

CDIM is a partitional and flat algorithm by nature. In the category of partitional (or flat) document clustering algorithms, we compare it with six popular methods. Four of these methods are K-means variants, one is based on Non-Negative Matrix Factorization (NMF) [52], and one is based on spectral clustering (SC) [12]. In addition to the comparison with partitional and flat methods, we further highlight the quality of CDIM by comparing it with five famous hierarchical document clustering methods. These hierarchical methods include frequent items based methods FIHC [18] and HFTC [6], NMF based method Rank-2 NMF [34], agglomerative method UPGMA and bisecting method Bi-Kmeans [30]. Note that Bi-Kmeans under hierarchical methods is same as the RB-I2 K-means variant under partitional methods.

The four K-means variants are selected from the CLUTO Toolkit [32] based on their strong performances reported in the literature [47,49]. Two of them are direct K-way clustering methods while the remaining two are repeated bisection methods that obtain K clusters by repeatedly performing two-way partitioning of clusters. For each of these two types of methods, we consider two different objective functions. One objective function maximizes the sum of similarities between documents and their cluster mean. The direct and repeated bisection methods that use this objective function are identified as Direct-I2 and RB-I2, respectively. The second objective function that we consider maximizes the ratio of I2 and E1, where I2 is the intrinsic (based on cluster cohesion) objective function defined above and E1 is an extrinsic (based on separation) function that minimizes the sum of the normalized pairwise similarities of documents within clusters with the rest of the documents. The direct and repeated bisection methods that use this hybrid objective function are identified as Direct-H2 and RB-H2, respectively.

For NMF, we use the implementation provided in the DTU:Toolbox.[6] Specifically, we use the multiplicative update rule with Euclidean measure for approximating the term-document matrix.

For spectral clustering (SC), we use the publicly available implementation.[7] We use the Nystrom method without orthogonalization because it is more scalable and performance difference with other spectral clustering methods available in the tool is not significant. After some initial experimentation on input values, the number of samples is set equal to the size of the corpus and sigma is set equal to 50 based on better performance.

[1] http://labs-repos.iit.demokritos.gr/skel/i-config/downloads/.

[2] http://www.cs.cornell.edu/People/pabo/movie-review-data/.

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

[4] http://trec.nist.gov.

[5] http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html.

[6] http://cogsys.imm.dtu.dk/toolbox/.

[7] http://alumni.cs.ucsb.edu/wychen/sc.html.

Table 1
Data sets and their characteristics.

#	Name	Documents (N)	Terms (M)	Categories (K)
1	pu	672	19 868	2
2	movie	1200	38 408	2
3	classic	7094	41 681	4
4	reviews	4069	23 220	5
5	hitech	2301	13 170	6
6	tr31	927	10 128	7
7	tr41	878	7454	10
8	ohscal	11 162	11 465	10
9	re0	1504	2886	13
10	wap	1560	8460	20
11	re1	1657	3758	25
12	reuters	8293	18 933	65

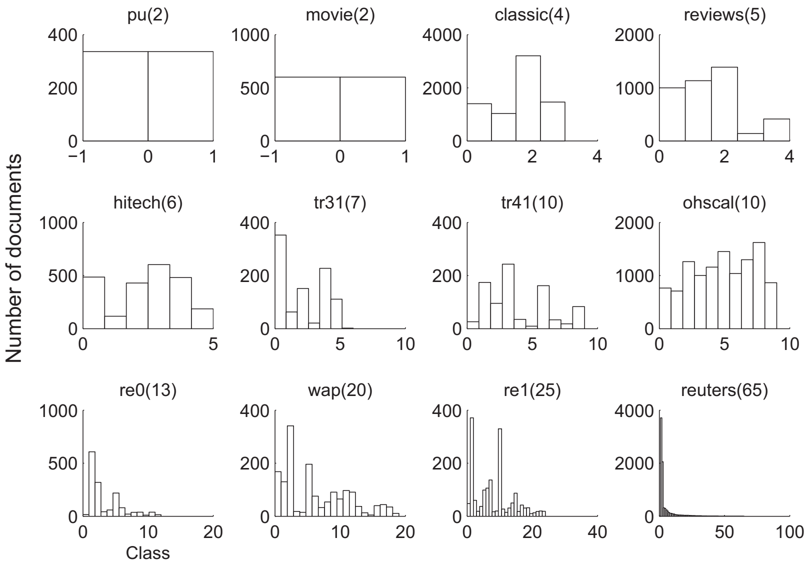


Fig. 2. Class distribution of the twelve text data sets is shown in ascending order of number of classes. Title of each subplot is in *dataset-name(number-of-classes)* format.

In using the four K-means variants, the term-document matrix is defined by term-frequency-inverse-document-frequency (TF-IDF) values and the cosine similarity measure is adopted for document comparisons. The term-document matrix is defined by term frequency values for NMF and SC. Each method is run 10 times, every time starting with a random initialization of the clusters, and results are reported as average ± standard deviation of the performance measures.

In hierarchical methods, FIHC [18] is based on the idea of frequent itemset mining. Instead of computing pair-wise similarity of documents, cohesiveness of clusters is measured using frequent item sets. The second hierarchical method HFTC [6] measures the mutual overlap of frequent sets with respect to the sets of supporting documents to produce clusters based on frequent term sets. The published results of both of these methods and bisecting K-means [30] are borrowed from [18] for the sake of comparison. Missing results of UPGMA in [18] are computed and rest of the results are just copied. Rank-2 NMF [34] is an efficient NMF based document clustering method that recursively splits a candidate cluster to generate a hierarchy of document clusters. SmallK [7] software package is used for the computation of Rank-2 NMF clustering results. UPGMA (Un-weighted Pair Group Method with Arithmetic Mean) results are computed using the CLUTO Toolkit [32].

5.3. Clustering validation measures

5.3.1. BCubed F-measure

The validation measure that we use is the BCubed metric [5,23]. In [1], Amigó et al. evaluate several extrinsic clustering validation measures both empirically and theoretically. They find that the BCubed precision and recall are the only measures that satisfy all desirable constraints for a good measure for clustering validation. The formal constraints that BCubed satisfies include cluster homogeneity, cluster completeness, rag bag and clusters size versus quantity. Cluster homogeneity means that clusters should not mix items belonging to different categories. Cluster completeness means that all items of same category should be grouped in the same cluster. Rag bag means that there should be a miscellaneous type of cluster to contain items that do not fall clearly in any cluster. Clusters size versus quantity means that a small error in a big cluster should be preferable to a large number of small errors in small clusters. BCubed is defined as below.

Let $L(o)$ and $C(o)$ be the category and cluster of an object o. Then, the correctness of the relation between objects o and o' in the clustering can be defined as

$$Correct(o,o') = \begin{cases} 1 & \text{iff } L(o) = L(o') \leftrightarrow C(o) = C(o') \\ 0 & \text{otherwise} \end{cases}.$$

This means that the relationship between two objects that share a category is correct if and only if they are in the same cluster. BCubed precision (BP) and BCubed recall (BR) can now be defined as follows:

$$BP = Avg_o[Avg_{o'.C(o)=C(o')}[Correct(o,o')]]$$

$$BR = Avg_o[Avg_{o'.L(o)=L(o')}[Correct(o,o')]]$$

BCubed precision and recall are computed for each data object or document, illustrated in Fig. 3. To obtain a single evaluation value, BCubed precision and recall are combined using the harmonic mean formula:

$$BF = 2 \times \frac{BP \times BR}{BP + BR} \tag{23}$$

The BCubed F-measure (BF) ranges from 0 to 1 with larger value signifying better clustering.

5.3.2. F-measure

We compute standard F-measure value also to compare CDIM's performance with results of HFTC, UPGMA and FIHC published in [18].

Let L_i and C_j represent the ith category and jth cluster, then F-measure $F(L_i, C_j)$ is computed as below:

$$Precision(L_i, C_j) = \frac{n_{ij}}{C_j}$$
$$Recall(L_i, C_j) = \frac{n_{ij}}{L_i}$$
$$F(L_i, C_j) = 2 \times \frac{Precision(L_i, C_j) \times Recall(L_i, C_j)}{Precision(L_i, C_j) + Recall(L_i, C_j)}$$

where n_{ij} is the number of members of category L_i in cluster C_j. Overall F-measure $F(C)$ quantifies the quality of a complete clustering solution C as:

$$F(C) = \sum_{L_i \in L} \frac{|L_i|}{|D|} \max_{C_j \in C} F(L_i, C_j) \tag{24}$$

where L denotes all categories; C denotes all clusters; $|L_i|$ denotes the number of documents in category L_i; and $|D|$ denotes the total number of documents in the data set. $F(C)$ ranges from 0 to 1, and larger value indicates a higher accuracy of clustering.

6. Results and discussion

In this section, we present the results of our experimental evaluations. These are divided into clustering quality comparison, cluster understanding and visualization, some additional experiments, and discussion.

6.1. Clustering quality

A common approach to evaluating clustering methods is through clustering quality analysis. We compare the clusterings produced by CDIM to that produced by ten popular methods on twelve public and standard text data sets. Since CDIM is an iterative partitioning method, its primary competitors are partitional and flat methods. In the following subsection we first justify the superiority of CDIM by comparing it with partitional or flat methods, then we compare CDIM with popular hierarchical document clustering methods in the next subsection.

6.1.1. CDIM versus partitional or flat document clustering methods

In this experimental setup, the desired number of clusters K for each data set is set equal to the number of categories in that data set (see Table 1). We report performance with BCubed F-measure, averaged over ten generated clusterings starting with random initial partitions.

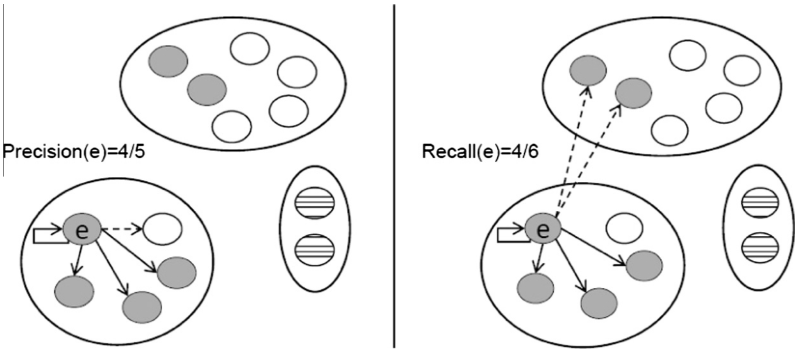


Fig. 3. Example of computing the BCubed precision and recall for one document [1].

Table 2 compares the clustering quality of CDIM using the Relative Risk (RR) and the measurement of discrimination information (MDI) measures with the other six methods. The highest average performance for each data set is highlighted in bold.

The two CDIM variants achieve highest score in nine out of twelve data sets altogether, Table 2. Looking into their individual performances, CDIM using relative risk i.e. CDIM-RR outperforms its six competitors (excluding CDIM-MDI) in seven out of twelve data sets. The pattern of consistently better performance for smaller values of K is also visible. This is attributable to the lesser resolution power of the multi-way comparisons ($\hat{d}_{ik} = d_{ik} - \bar{d}_{ik}, \forall k$) that are required in CDIM-RR for document assignment. Interestingly, CDIM-MDI remains unbeaten by the six competitors (excluding CDIM-RR) in eight out of twelve data sets and its performance looks enhancing with increasing values of K. The incorporation of the log factor with priors based normalization, as in Eqs. (4) and (5), causes this improvement.

Investigating the details of the data sets gives a better picture of the winning conditions for the algorithms. Basic data statistics: average terms per document, average terms per category and average documents per category for each data set are reported in Table 4. It is notable that for data sets having larger values of average number of terms per document e.g. values greater than ten, CDIM-RR outperforms others whereas for data sets having smaller values of average number of terms per document e.g. values less than or equal to six, CDIM-MDI outperforms others. The data sets where some other algorithm wins i.e. ohscal and tr41, the winning margin is not very high and the winner has score value within one standard deviation of CDIM. This observation about results based on the number of average terms per document is shown in Fig. 4. Hence we find that for large documents, the simpler measure i.e. relative risk proves to be more powerful discrimination measure as compared to MDI. If documents in the data set are short than MDI becomes more discriminative due to the use of prior probabilities and log factor calculations.

Table 3 presents the comparison of CDIM variants with each other. The columns give BCubed values of CDIM using relative risk, MDI, repeated bisection using relative risk and MDI, and MDI using symmetric priors, in order. The comparison shows that CDIM-MDI and CDIM-RR win most of the times. CDIM-MDI-S also gives reasonable results. CDIM-DRDC performs great on smaller values of K but its performance degrades on data sets with larger K values. Based on this evaluation of CDIM variants, we pick CDIM-MDI to represent CDIM in our further comparisons with hierarchical clustering methods.

6.1.2. CDIM versus hierarchical document clustering methods

To compare CDIM with hierarchical methods, we borrow results of FIHC, HFTC, UPGMA, and bisecting K-means from [18]. We extend the results table (Table 7 in [18]) by filling in the missing results of UPGMA for classic and reuters data sets; and including results of CDIM and Rank-2 NMF. Table 5 presents the comparison of CDIM with these hierarchical methods. The comparison contains five data sets including classic, hitech, re0, reuters and wap. The description of data sets has already been provided in Section 5.1. Four settings are evaluated for each data set based on the number of clusters which is set to be 3, 15, 30 or 60. Average F-measure value of a method for each data set is presented. Overall average is also calculated in the last row. The highest F-measure value for a setting is highlighted in bold typeface. Highest average F-measure value for a data set over all settings is highlighted in bold typeface and a $*$ mark. It is evident from the results (Table 5) that CDIM achieves the highest value of overall F-measure averaged over all data sets. Comparing the performance on individual data sets, if we check the average F-measure value for each data set; CDIM wins in three out of five while UPGMA wins in two out of five data sets. Performance of CDIM is most of the times better than its competitors when number of clusters is set somewhere near to the actual categories in the data set. Hierarchical methods show good performance on reuters data having 65 categories. One reason is the highly skewed distribution of this data set (Fig. 2), that naturally favors the hierarchical methods that make large cluster and obtain high recall value.

6.1.3. Statistical significance tests

In order to validate our results we apply Friedman's test, a non-parametric test popularly used for significance testing of multiple algorithms on multiple data sets [15]. The test is applied using 0.05 level of significance. Performance of CDIM-MDI

Table 2
Comparison of clustering quality of CDIM via the BCubed F-measure. Averages and standard deviations obtained from 10 runs are shown.

Data	CDIM-RR	CDIM-MDI	Direct-I2	Direct-H2	RB-I2	RB-H2	NMF	SC
pu	**0.732 ± .08**	0.582 ± .03	0.565 ± .02	0.553 ± .02	0.565 ± .02	0.553 ± .02	0.612 ± .04	0.628 ± .03
movie	**0.556 ± .02**	0.533 ± .02	0.533 ± .02	0.522 ± .01	0.533 ± .02	0.522 ± .01	0.510 ± .01	0.519 ± .01
classic	0.668 ± .08	0.687 ± .04	0.669 ± .04	0.675 ± .04	0.671 ± .03	0.680 ± .02	0.512 ± .03	**0.717 ± .01**
reviews	0.675 ± .07	**0.732 ± .07**	0.627 ± .06	0.626 ± .06	0.609 ± .04	0.669 ± .03	0.552 ± .03	0.425 ± .00
hitech	0.442 ± .03	**0.500 ± .02**	0.391 ± .02	0.380 ± .02	0.394 ± .02	0.390 ± .03	0.399 ± .02	0.408 ± .01
tr31	**0.613 ± .09**	0.594 ± .06	0.585 ± .05	0.575 ± .05	0.553 ± .07	0.572 ± .05	0.362 ± .03	0.443 ± .01
tr41	0.578 ± .06	0.508 ± .03	**0.608 ± .02**	0.584 ± .03	0.602 ± .05	0.590 ± .04	0.361 ± .04	0.310 ± .00
ohscal	0.428 ± .04	0.309 ± .03	0.422 ± .02	0.417 ± .03	**0.432 ± .01**	0.427 ± .01	0.250 ± .02	0.300 ± .01
re0	0.411 ± .02	**0.431 ± .02**	0.382 ± .02	0.382 ± .01	0.397 ± .03	0.375 ± .01	0.345 ± .02	0.323 ± .01
wap	0.445 ± .02	**0.478 ± .03**	0.462 ± .01	0.444 ± .01	0.465 ± .02	0.438 ± .02	0.299 ± .02	0.411 ± .01
re1	0.393 ± .04	**0.456 ± .03**	0.443 ± .02	0.436 ± .02	0.416 ± .01	0.418 ± .03	0.301 ± .03	0.299 ± .01
reuters	0.349 ± .04	**0.362 ± .07**	0.274 ± .02	0.283 ± .03	0.275 ± .02	0.268 ± .03	0.280 ± .03	0.219 ± .01

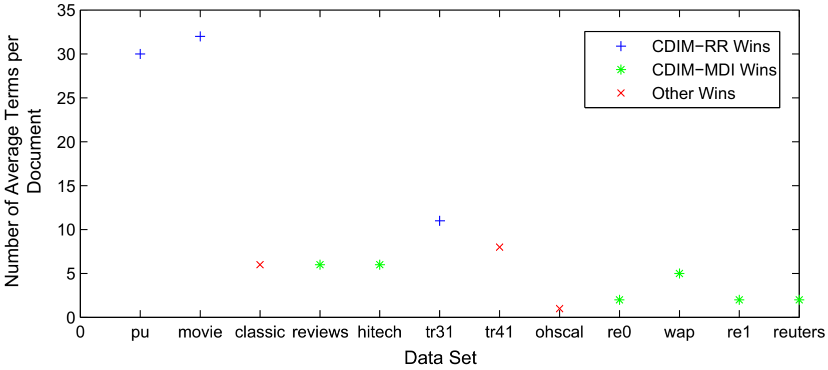


Fig. 4. Terms per document value can be used to select between RR and MDI.

Table 3
Comparison of clustering quality among CDIM variants via the BCubed F-measure. Averages and standard deviations obtained from 10 runs are shown.

Data	CDIM-RR	CDIM-MDI	CDIM-RR-RB	CDIM-MDI-RB	CDIM-DRDC	CDIM-MDI-S
pu	**0.732 ± .08**	0.582 ± .03	**0.732 ± .08**	0.582 ± .03	0.687 ± .08	0.582 ± .03
movie	0.556 ± .02	0.533 ± .02	0.556 ± .02	0.533 ± .02	0.562 ± .04	**0.572 ± .05**
classic	0.633 ± .03	0.687 ± .04	0.500 ± .05	0.509 ± .03	0.704 ± .07	**0.743 ± .05**
reviews	0.675 ± .07	**0.732 ± .07**	0.672 ± .03	0.674 ± .03	0.632 ± .05	0.646 ± .05
hitech	0.442 ± .03	**0.500 ± .02**	0.386 ± .03	0.431 ± .03	0.338 ± .02	0.462 ± .02
tr31	**0.613 ± .09**	0.594 ± .06	0.491 ± .06	0.465 ± .03	0.558 ± .08	0.590 ± .08
tr41	**0.578 ± .06**	0.508 ± .03	0.514 ± .06	0.517 ± .03	0.535 ± .05	0.536 ± .03
ohscal	**0.428 ± .04**	0.309 ± .03	0.375 ± .02	0.283 ± .02	0.351 ± .02	0.318 ± .02
re0	0.411 ± .02	**0.431 ± .02**	0.401 ± .01	0.419 ± .02	0.399 ± .01	0.390 ± .01
wap	0.445 ± .02	**0.478 ± .03**	0.395 ± .03	0.363 ± .02	0.396 ± .02	0.427 ± .02
re1	0.393 ± .04	**0.456 ± .03**	0.318 ± .03	0.334 ± .01	0.292 ± .02	0.391 ± .01
reuters	0.349 ± .04	0.362 ± .07	0.249 ± .02	0.218 ± .01	0.265 ± .01	**0.467 ± .02**

is found to be significantly better than all of the partitional or flat clustering methods under comparison. When compared to hierarchical methods, CDIM-MDI achieves top position according to mean rank value. Performance of CDIM-MDI is significantly better than Bi Kmeans, HFTC and Rank-2 NMF. Although CDIM gets the highest mean rank value, yet the difference from FIHC and UPGMA is not statistically significant. However superior performance of CDIM over FIHC and UPGMA in providing better cluster understanding is demonstrated in next section (Section 6.2).

We tested also the significance of difference between the results of CDIM variants using Friedman's test. CDIM-RR and CDIM-MDI achieve top positions according to the mean rank. The performance difference of CDIM-RR and CDIM-MDI is not statistically significant from each other.

6.2. Cluster understanding and visualization

A key application of data clustering is corpus understanding. In the case of document clustering, it is important that clustering methods output information that can readily be used to interpret the clusters and their documents. CDIM is based on term discrimination information and each of its cluster is describable by the highly discriminating terms in it.

Table 4
Data sets statistics (rounded off values). Terms (M), Documents (N), Categories (K).

#	Data	M/N	M/K	N/K	Winner
1	pu	30	9934	336	CDIM-RR
2	movie	32	19,204	600	CDIM-RR
3	classic	6	10,420	1774	SC
4	reviews	6	4644	814	CDIM-MDI
5	hitech	6	2195	384	CDIM-MDI
6	tr31	11	1447	132	CDIM-RR
7	tr41	8	745	88	Direct-I2
8	ohscal	1	1147	1116	RB-I2
9	re0	2	222	116	CDIM-MDI
10	wap	5	423	78	CDIM-MDI
11	re1	2	150	66	CDIM-MDI
12	reuters	2	291	128	CDIM-MDI

Table 5
Comparison of clustering quality of CDIM via the Overall F-measure. Highest value has bold typeface. Highest average value has a $*$ mark and bold typeface.

Data	No. of clusters	CDIM-MDI	FIHC	UPGMA	Bi-Kmeans	HFTC	Rank-2-NMF
	3	**0.63**	0.62	0.45	0.59	n/a	0.58
classic	15	0.66	0.52	**0.72**	0.46	n/a	0.42
(4)	30	0.55	0.52	**0.73**	0.43	n/a	0.39
	60	0.46	0.51	**0.67**	0.27	n/a	0.23
	Avg.	0.58	0.54	**0.64***	0.44	0.61	0.41
	3	**0.55**	0.45	0.33	0.54	n/a	0.48
hitech	15	**0.52**	0.42	0.33	0.44	n/a	0.45
(6)	30	0.46	0.41	**0.47**	0.29	n/a	0.33
	60	**0.47**	0.41	0.40	0.21	n/a	0.23
	Avg.	**0.50***	0.42	0.38	0.37	0.37	0.37
	3	0.48	**0.53**	0.36	0.34	n/a	0.40
re0	15	**0.48**	0.45	0.47	0.38	n/a	0.39
(13)	30	**0.47**	0.43	0.42	0.38	n/a	0.32
	60	**0.40**	0.38	0.34	0.28	n/a	0.32
	Avg.	**0.46***	0.45	0.40	0.34	0.43	0.36
	3	0.49	**0.58**	0.49	0.48	n/a	0.57
reuters	15	0.58	0.61	**0.67**	0.42	n/a	0.60
(65)	30	0.54	0.61	**0.70**	0.35	n/a	0.56
	60	0.52	0.60	**0.63**	0.30	n/a	0.45
	Avg.	0.53	0.60	**0.62***	0.39	0.49	0.55
	3	0.39	**0.40**	0.39	**0.40**	n/a	0.36
wap	15	**0.60**	0.56	0.49	0.57	n/a	0.25
(20)	30	**0.61**	0.57	0.58	0.44	n/a	0.19
	60	0.53	0.55	**0.59**	0.37	n/a	0.20
	Avg.	**0.53***	0.52	0.51	0.45	0.35	0.25
Overall	Average	**0.52**	0.51	0.51	0.40	0.45	0.39

We illustrate the understanding provided by CDIM's output by comparing the top 10 most discriminating terms (stemmed words) for each cluster of the ohscal data set in Table 6. The ohscal data set contains publications from 10 different medical subject areas (antibodies, carcinoma, DNA, in vitro, molecular sequence data, pregnancy, prognosis, receptors, risk factors, and tomography). The first part (top) of the table shows the top 10 terms in each class using TF-IDF weights. The second part demonstrates the top 10 most discriminating terms provided by our approach, CDIM. The corresponding class name is presented with the cluster number for the sake of easier comparison. Identification of the corresponding class is done by comparing the CDIM top terms with the class top terms manually. The third part of the table shows the top 10 most descriptive terms obtained by analyzing the clusters generated using UPGMA. These are the terms that contribute the most to the average similarity between the documents of each cluster. The terms are obtained using CLUTO toolkit. The last part of the table shows the top 10 (if available) descriptions/labels of the (sub) clusters produced by FIHC. The cluster labels provided by FIHC are presented beside the cluster number.

In case of CDIM only, it is easy to determine the category of most clusters by looking at the top ten terms: cluster 2 = carcinoma, cluster 3 = antibodies, cluster 4 = prognosis, cluster 5 = pregnancy, cluster 6 = risk factors, cluster 7 = DNA, cluster 9 = receptors, cluster 10 = tomography. The categories molecular sequence data and in vitro do not appear to have a well-defined cluster; molecular sequence data has some overlap with cluster 7 while in vitro has some overlap with clusters 1 and 9. Nonetheless, clusters 1 and 8 still give coherent meaning to the documents they contain. For UPGMA, there is a

Table 6
Comparison of top 10 terms (stemmed words) in ohscal data set using TF-IDF, CDIM, UPGMA and FIHC.

Class	Top 10 terms in class (TF-IDF)
DNA	‘0’, ‘cell’, ‘gene’, ‘patient’, ‘hepat’, ‘tumor’, ‘chromosom’, ‘viru’, ‘human’, ‘protein’
Carcinoma	‘0’, ‘tumor’, ‘patient’, ‘cell’, ‘cancer’, ‘respons’, ‘surviv’, ‘renal’, ‘dose’, ‘squamou’
Tomography	‘ct’, ‘imag’, ‘0’, ‘patient’, ‘tumor’, ‘diseas’, ‘scan’, ‘lesion’, ‘arteri’, ‘injuri’
In-Vitro	‘il’, ‘0’, ‘cell’, ‘activ’, ‘protein’, ‘rate’, ‘releas’, ‘alpha’, ‘respons’, ‘stimul’
Antibodies	‘0’, ‘cell’, ‘patient’, ‘infect’, ‘anti’, ‘vaccin’, ‘antigen’, ‘tumor’, ‘activ’, ‘protein’
Risk-Factors	‘patient’, ‘infect’, ‘cancer’, ‘hiv’, ‘diseas’, ‘coronari’, ‘women’, ‘children’, ‘factor’, ‘0’
Prognosis	‘0’, ‘surviv’, ‘tumor’, ‘diseas’, ‘cell’, ‘treatment’, ‘stage’, ‘month’, ‘ventricular’, ‘eye’
Receptors	‘0’, ‘cell’, ‘il’, ‘beta’, ‘alpha’, ‘patient’, ‘bind’, ‘rate’, ‘insulin’, ‘express’
Pregnancy	‘0’, ‘fetal’, ‘pregnanc’, ‘patient’, ‘women’, ‘rate’, ‘matern’, ‘infant’, ‘level’, ‘gestat’
Mol-Seq-Dat	‘0’, ‘gene’, ‘cell’, ‘beta’, ‘peptid’, ‘protein’, ‘bind’, ‘alpha’, ‘factor’, ‘express’
k	Top 10 discriminative terms for cluster *k* (CDIM)
1 (Mol.)	‘platelet’, ‘kg’, ‘mg’, ‘dose’, ‘min’, ‘plasma’, ‘pressur’, ‘flow’, ‘microgram’, ‘antagonist’
2 (Carc.)	‘carcinoma’, ‘tumor’, ‘cancer’, ‘surviv’, ‘chemotherapi’, ‘stage’, ‘recurr’, ‘malign’, ‘resect’, ‘therapi’
3 (Anti.)	‘antibodi’, ‘antigen’, ‘viru’, ‘anti’, ‘infect’, ‘hiv’, ‘monoclon’, ‘ig’, ‘immun’, ‘sera’
4 (Prog.)	‘patient’, ‘complic’, ‘surgeri’, ‘ventricular’, ‘infarct’, ‘oper’, ‘eye’, ‘coronari’, ‘cardiac’, ‘morta’
5 (Preg.)	‘pregnanc’, ‘fetal’, ‘gestat’, ‘matern’, ‘women’, ‘infant’, ‘deliveri’, ‘birth’, ‘labor’, ‘pregnant’
6 (Risk.)	‘risk’, ‘alcohol’, ‘age’, ‘children’, ‘cholesterol’, ‘health’, ‘factor’, ‘women’, ‘preval’, ‘popul’
7 (DNA)	‘gene’, ‘sequenc’, ‘dna’, ‘mutat’, ‘protein’, ‘chromosom’, ‘transcript’, ‘rna’, ‘amino’, ‘structur’
8 (In-V.)	‘contract’, ‘muscle’, ‘relax’, ‘microm’, ‘calcium’, ‘effect’, ‘respons’, ‘antagonist’, ‘releas’, ‘action’
9 (Rece.)	‘il’, ‘receptor’, ‘cell’, ‘stimul’, ‘bind’, ‘growth’, ‘gamma’, ‘alpha’, ‘insulin’, ‘0’
10 (Tomo.)	‘ct’, ‘imag’, ‘comput’, ‘tomographi’, ‘scan’, ‘lesion’, ‘magnet’, ‘reson’, ‘cerebr’, ‘tomograph’
k	Top 10 descriptive terms for cluster *k* (UPGMA)
1	‘flap’, ‘hernia’, ‘reconstruct’, ‘lip’, ‘diaphragmat’, ‘eyelid’, ‘glove’, ‘defect’, ‘repair’, ‘abdomin’
2	‘magnesium’, ‘sulfat’, ‘cartilag’, ‘dermat’, ‘eclampsia’, ‘proteoglycan’, ‘terbutalin’, ‘seizur’, ‘0’, ‘cell’
3 (Prog.)	‘eye’, ‘laser’, ‘retin’, ‘visual’, ‘intraocular’, ‘iop’, ‘yag’, ‘glaucoma’, ‘cataract’, ‘detach’
4	‘asa’, ‘claim’, ‘readmiss’, ‘payment’, ‘exceed, ‘kidnei’, ‘iii’, ‘arylsulfatas’, ‘anaesthesia’, ‘clos’
5 (DNA)	‘0’, ‘cell’, ‘protein’, ‘gene’, ‘receptor’, ‘antibodi’, ‘il’, ‘infect’, ‘activ’, ‘hiv’
6	‘sutur’, ‘adher’, ‘rad’, ‘ccr’, ‘mbq’, ‘moab’, ‘mci’, ‘keratoplasti’, ‘bar’, ‘perin’
7	‘celiac’, ‘oligom’, ‘agglutin’, ‘gliadin’, ‘mannan’, ‘thermodynam’, ‘cell’, ‘peptid’, ‘cereal’, ‘axial’
8	‘syndrom’, ‘neurolept’, ‘equina’, ‘cauda’, ‘exchang’, ‘marfan’,‘dyskinesia’, ‘ankylos’, ‘0’, ‘guillain’
9 (Preg.)	‘patient’, ‘risk’, ‘fetal, ‘arteri’, ‘pregnanc’, women’, coronari’, rate’, heart’, ‘diabet’
10 (Carc.)	‘patient’, ‘tumor’, ‘carcinoma’, ‘cancer’, ‘ct, ‘cell’, ‘diseas’, ‘surviv’, ‘imag’, ‘lesion’
k	Top 10 descriptions for cluster *k* (FIHC)
1 (ag)	‘activ’, ‘factor’, ‘ag’, ‘associ’, ‘increa’, ‘patient’, ‘result’, ‘risk’, ‘studi’, ‘disea’
2 (antibodi)	‘antibodi’,‘cell’, ‘patient’, ‘result’, ‘specif’, ‘studi’, ‘infect’
3 (cell)	‘activ’,‘cell’, ‘effect’, ‘function’, ‘human’, ‘increa’, ‘induc’, ‘inhibit’, ‘level’, ‘patient’
4 (dose)	‘dose’, ‘effect’, ‘increa’, ‘result’, ‘time’, ‘treatment’, ‘us’
5 (gene)	‘activ’, ‘protein’, ‘cell’, ‘gene’, ‘result’, ‘specif’, ‘studi’, ‘suggest’
6 (increa)	‘determin’, ‘effect’,‘increa’, ‘measur’, ‘normal’, ‘observ’, ‘patient’, ‘respon’, ‘risk’, ‘signif’
7 (independ)	‘independ’
8 (patient)	‘five’, ‘patient’, ‘follow’, ‘result’, ‘studi’, ‘initi’, ‘mean’,‘month’, ‘rate’, ‘respon’
9 (pregnanc)	‘pregnanc’
10 (serum)	‘effect’, ‘level’, ‘increase’, ‘patient’, ‘result’, ‘suggest’

mix of terms from many categories for almost all clusters. Four out of ten clusters could be identified to represent the categories mentioned beside their cluster number. The descriptions provided by FIHC are also not as understandable as the terms provided by CDIM. Enough number of descriptions are not readily available many times. Note that CDIM can also generate the label or topic of a cluster like FIHC by choosing the top one or three most discriminating terms. Thus CDIM produces better discriminative, representative and understandable terms for clusters than UPGMA and FIHC.

Since CDIM finds clusters in a K-dimensional discrimination information space, the distribution of documents among clusters can be visualized via simple scatter plots. The 2-dimensional scatter plot of documents in the pu data set is shown in Fig. 5 for clustering produced by CDIM-RR. The x- and y-axes in this figure correspond to document discrimination information for cluster 1 and 2 (d_{i1} and d_{i2}), respectively. When $d_{i1} > d_{i2}$ then document $\mathbf{x}_i$ belongs to cluster 1, and vice versa. Thus, the two clusters are spread along the two axes starting from the origin. For illustration purposes, the true labels of the documents are also shown in Fig. 5, indicating that a high quality clustering is obtained.

Such scatter plots can be viewed for any pair of clusters when $K > 2$. Since CDIM's document assignment decision is based upon document discrimination scores ($\hat{d}_{ik}, \forall k$), scatter plots of documents in this space are also informative; each axis quantifies how relevant a document is to a cluster in comparison to the remaining clusters.

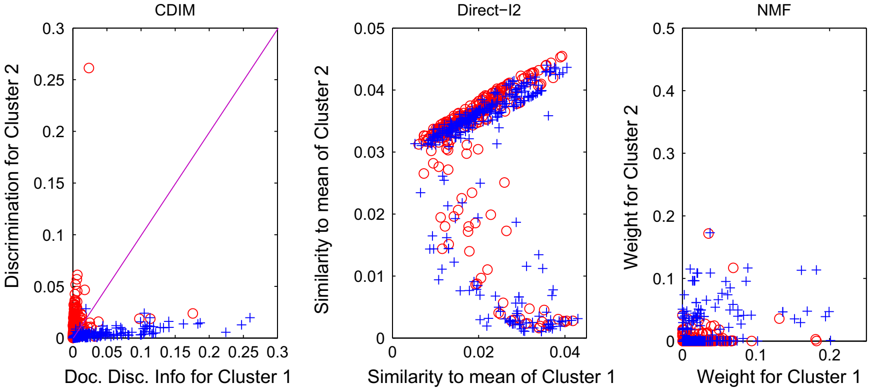


Fig. 5. Scatter plot of documents in pu data set projected onto the 2-D discrimination information space (CDIM-RR), similarity to cluster mean space (Direct-I2), and weight space (NMF). True labels are indicated by different color markers. CDIM has the best document projection and separation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

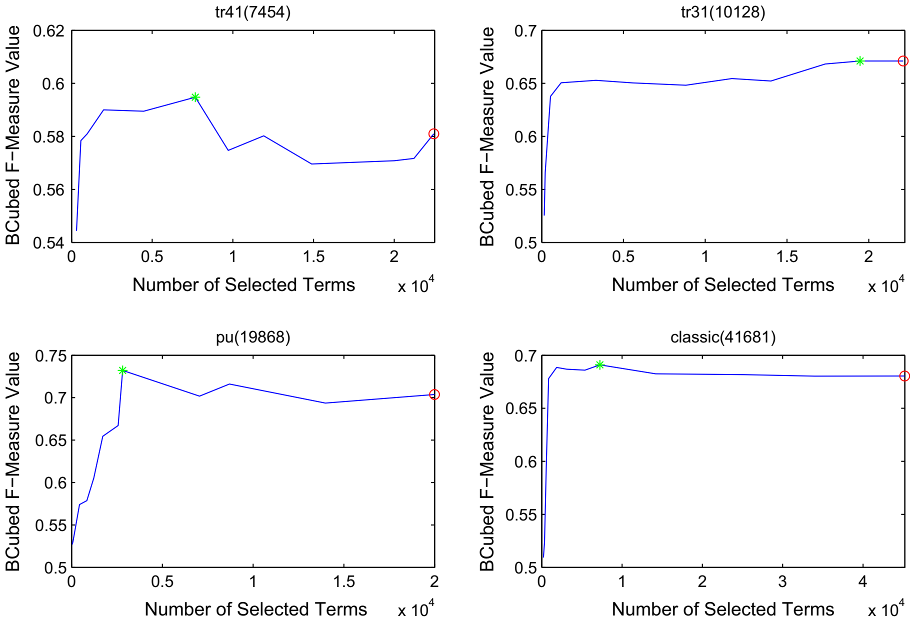


Fig. 6. Impact of term selection on clustering quality. The 'green $*$' and the 'red o' markers on the line show the best BCubed F-measure value and the BCubed F-measure value by using all terms respectively. Sub-plot title is in *dataset-name(number-of-terms in the data set)* format. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.3. Additional experiments and discussion

We present some additional results and provide a discussion of some implementation issues in this section.

6.3.1. Term discrimination information distribution and term selection

Typically, the distribution of term discrimination information has a long narrow tail. In other words, a small fraction of terms provide high discriminative power while the vast majority provide little discriminative power. This observation suggests that term selection via the parameter t can reduce the space complexity of CDIM without impacting clustering quality significantly. Except for the experiments reported in this subsection, all other results are obtained with $t = 0$.

We evaluate the impact of term selection by comparing clustering quality with increasing value of term selection parameter t. Fig. 6 shows the variation of clustering quality using CDIM-RR with sum of significant terms for all clusters ($\sum_{i=1}^{K}|T_i|$) in the data set. It is seen that clustering quality does not degrade much with significant decrease in number of terms. In fact, an improve in quality is observed through term selection many times. For example, when the number of terms is reduced to

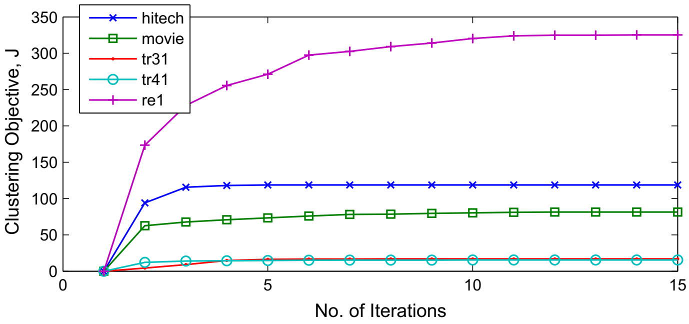


Fig. 7. Convergence curves of CDIM.

2,816 from 19,868 in the pu data set, the clustering quality actually improves. These results show that CDIM is scalable and robust to term selection.

6.3.2. Convergence

In practice, smooth and rapid convergence of algorithms is highly desirable. CDIM maximizes its objective function by using a two-step greedy procedure that ensures that the objective function is non-decreasing from one iteration to the next. This procedure corresponds to the naive semi-supervised learning approach of labeling and learning, which has also been shown to be convergent [54].

Fig. 7 shows the convergence curves of CDIM on five data sets. This figure demonstrate the smooth convergence behavior of CDIM, and highlights the fact that convergence is achievable within 15 iterations. CDIM is able to find a local optimum solution only, dependent on the initial partitioning chosen. This characteristic, however, is present in many partitional or flat clustering methods like the *K*-means and NMF algorithms.

7. Conclusion and future work

In this paper, we propose and evaluate a document clustering framework, CDIM. CDIM finds clusters in a *K*-dimensional space in which documents are well discriminated. It does this by maximizing the sum of the discrimination information provided by documents for their clusters minus that provided for the remaining clusters. Document discrimination information is computed from the discrimination information provided by the terms in it. Term discrimination information can be estimated from the document collection via its relative risk or any other discrimination measure e.g. MDI, etc. An advantage of using a measure of discrimination information is that it also quantifies the degree of relatedness of a term to its context in the collection. Thus, CDIM produces clusters that are readily interpretable by their highly discriminating terms.

We conduct extensive experimental evaluations of CDIM. We compare its cluster quality with that of ten popular clustering methods on twelve data sets. CDIM significantly outperforms popular partitioning, NMF-based and spectral clustering methods and its performance is on par with the best hierarchical document clustering methods. The superior understanding provided by CDIM's output is also demonstrated, enabling documents in the clusters to be identifiable as belonging to specific contexts or topics.

Our results suggest that CDIM is a practically useful document clustering method. Its core idea of clustering in spaces defined by corpus-based discrimination or relatedness information holds much potential for future extensions and improvements. In addition to implementing soft-CDIM, the soft clustering version of CDIM, we would like to investigate other measures of discrimination/relatedness information, extend and evaluate CDIM for constraint based clustering, and co-clustering. Use of external sources like WordNet and Wikipedia with CDIM, CDIM with intelligent seeding, CDIM using item sets and sequences are also some promising future extensions.

Acknowledgements

This work is in part supported by ICT R&D program of MSIP/IITP [B0101-15-0525], South Korea; the Department of Computer Science at LUMS, Pakistan; and Higher Education Commission of Pakistan.

References

[1] E. Amigó, J. Gonzalo, J. Artiles, F. Verdejo, A comparison of extrinsic clustering evaluation metrics based on formal constraints, Inform. Retrieval 12 (4) (2009) 461–486.

[2] Henry Anaya-Sánchez, Aurora Pons-Porrata, Rafael Berlanga-Llavori, A document clustering algorithm for discovering and describing topics, Pattern Recogn. Lett. 31 (6) (2010) 502–510.

[3] N.O. Andrews, E.A. Fox, Recent developments in document clustering, Technical report, Dept. of Computer Science, Virginia Tech, 2007.
[4] David Arthur, Sergei Vassilvitskii, k-means++: The advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.
[5] A. Bagga, B. Baldwin, Entity-based cross-document coreferencing using the vector space model, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, ACL, 1998, pp. 79–85.
[6] Florian Beil, Martin Ester, Xiaowei Xu, Frequent term-based text clustering, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 436–442.
[7] Richard Boyd, Barry Drake, Da Kuang, Haesun Park, Smallk is a C++/Python high-performance software library for nonnegative matrix factorization (nmf) and hierarchical and flat clustering using the nmf; current version 1.2.0, June 2014. <http://smallk.github.io/.
[8] D. Cai, CJ Van Rijsbergen, Learning semantic relatedness from term discrimination information, Expert Syst. Appl. 36 (2) (2009) 1860–1875.
[9] Deng Cai, Xiaofei He, Jiawei Han, Locally consistent concept factorization for document clustering, IEEE Trans. Knowl. Data Eng. 23 (6) (2011) 902–913.
[10] Di Cai, An information-theoretic foundation for the measurement of discrimination information, IEEE Trans. Knowl. Data Eng. 22 (9) (2010) 1262–1273.
[11] S.F. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, Comput. Speech Language 13 (4) (1999) 359–393.
[12] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, Edward Y. Chang, Parallel spectral clustering in distributed systems, IEEE Trans. Pattern Anal. Machine Intell. 33 (3) (2011) 568–586.
[13] Y.M. Chung, J.Y. Lee, A corpus-based approach to comparative evaluation of statistical term association measures, J. Am. Soc. Inform. Sci. Technol. 52 (4) (2001) 283–296.
[14] Fernando De la Torre, Takeo Kanade, Discriminative cluster analysis, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 241–248.
[15] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Machine Learn. Res. 7 (2006) 1–30.
[16] I.S. Dhillon, S. Mallela, D.S. Modha, Information-theoretic co-clustering, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 89–98.
[17] Chris Ding, Tao Li, Adaptive dimension reduction using discriminant analysis and k-means clustering, in: Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 521–528.
[18] Benjamin C.M. Fung, Ke Wang, Martin Ester, Hierarchical document clustering using frequent itemsets, SDM, vol. 3, SIAM, 2003, pp. 59–70.
[19] W.A. Gale, K.W. Church, What is wrong with adding one, Corpus-based research into language, 1994, pp. 189–198.
[20] William A. Gale, Geoffrey Sampson, Good-turing frequency estimation without tears, J. Quant. Linguist. 2 (3) (1995) 217–237.
[21] Reynaldo Gil-García, Aurora Pons-Porrata, Dynamic hierarchical algorithms for document clustering, Pattern Recogn. Lett. 31 (6) (2010) 469–477.
[22] M.A.K. Haliday, R. Hassan, Cohesion in English, Longman, London, UK, 1976.
[23] J. Han, M. Kamber, Data Mining Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, 2011.
[24] Malik Tahir Hassan, Asim Karim, Impact of behavior clustering on web surfer behavior prediction, J. Inform. Sci. Eng. 27 (6) (2011) 1855–1870.
[25] Malik Tahir Hassan, Asim Karim, Clustering and understanding documents via discrimination information maximization, in: Advances in Knowledge Discovery and Data Mining, LNCS, vol. 7301, Springer, 2012, pp. 566–577.
[26] D.A. Hsieh, C.F. Manski, D. McFadden, Estimation of response probabilities from augmented retrospective observations, J. Am. Stat. Assoc. 80 (391) (1985) 651–662.
[27] X. Hu, X. Zhang, C. Lu, E.K. Park, X. Zhou, Exploiting wikipedia as external knowledge for document clustering, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 389–396.
[28] A.K. Jain, Data clustering: 50 years beyond k-means, Pattern Recogn. Lett. 31 (8) (2010) 651–666.
[29] Anil Jain, M.N. Murty, Patrick Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.
[30] Anil K. Jain, Richard C. Dubes, Algorithms for clustering data, Prentice-Hall Inc., 1988.
[31] K.N. Junejo, A. Karim, A robust discriminative term weighting based linear discriminant method for text classification, in: Eighth IEEE International Conference on Data Mining, 2008, pp. 323–332.
[32] G. Karypis, CLUTO – a clustering toolkit, Technical report, Dept. of Computer Science, University of Minnesota, Minneapolis, 2002.
[33] Hans-Peter Kriegel, Peer Kröger, Arthur Zimek, Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering, ACM Trans. Knowl. Discov. Data 3 (2009). 1:1–1:58.
[34] Da Kuang, Haesun Park, Fast rank-2 nonnegative matrix factorization for hierarchical document clustering, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 739–747.
[35] M. LeBlanc, J. Crowley, Relative risk trees for censored survival data, Biometrics 48 (2) (1992) 411–425.
[36] H. Li, J. Li, L. Wong, M. Feng, Y.P. Tan, Relative risk and odds ratio: a data mining perspective, in: PODS '05: Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2005.
[37] J. Li, G. Liu, L. Wong, Mining statistically important equivalence classes and delta-discriminative emerging patterns, in: KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007.
[38] Yanjun Li, Soon M. Chung, John D. Holt, Text document clustering based on frequent word meaning sequences, Data Knowl. Eng. 64 (1) (2008) 381–404.
[39] Yinglong Ma, Yao Wang, Beihong Jin, A three-phase approach to document clustering based on topic significance degree, Expert Syst. Appl. 41 (18) (2014) 8203–8210.
[40] H. Malik, D. Fradkin, F. Moerchen, Single pass text classification by direct feature weighting, Knowl. Inform. Syst. (2010) 1–20.
[41] Jane Morris, Graeme Hirst, Non-classical lexical semantic relations, in: Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, Association for Computational Linguistics, 2004, pp. 46–51.
[42] Jamal A. Nasir, Iraklis Varlamis, Asim Karim, George Tsatsaronis, Semantic smoothing for text clustering, Knowl.-Based Syst. 54 (2013) 216–229.
[43] R. Navigli, P. Velardi, Learning domain ontologies from document warehouses and dedicated web sites, Comput. Linguist. 30 (2) (2004) 151–179.
[44] Fuchun Peng, Dale Schuurmans, Combining naive bayes and n-gram language models for text classification, in: Fabrizio Sebastiani (Ed.), Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 2633, Springer, Berlin Heidelberg, 2003, pp. 335–350.
[45] G. Salton, Some hierarchical models for automatic document retrieval, Am. Document. 14 (3) (1963) 213–222.
[46] Gerard Salton, Chung-Shu Yang, CLEMENT T Yu, A theory of term importance in automatic text analysis, J. Am. Soc. Inform. Sci. 26 (1) (1975) 33–44.
[47] Michael Steinbach, George Karypis, Vipin Kumar, et al., A comparison of document clustering techniques, in: KDD Workshop on Text Mining, vol. 400, Boston, 2000, pp. 525–526.
[48] Andrea Tagarelli, George Karypis, Document clustering: the next frontier, Data Clustering: Algorithms and Applications, 2013, pp. 305–338.
[49] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison Wesley, New York, 2006.
[50] Bin Tang, Michael Shepherd, Malcolm Heywood, Xiao Luo, Comparing dimension reduction techniques for document clustering, in: Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 3501, Springer, Berlin/Heidelberg, 2005, pp. 1–3.
[51] Wei Xu, Yihong Gong, Document clustering by concept factorization, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2004, pp. 202–209.
[52] Wei Xu, Xin Liu, Yihong Gong, Document clustering based on non-negative matrix factorization, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2003, pp. 267–273.
[53] Zheng Xu, Xiangfeng Luo, Lin Mei, Chuanping Hu, Measuring the semantic discrimination capability of association relations, Concurr. Comput.: Pract. Exp. 26 (2) (2014) 380–395.

[54] J.C. Xue, G.M. Weiss, Quantification and semi-supervised classification methods for handling changes in class distribution, in: KDD-09: Proceedings of the 15th Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 897–906.

[55] Taiping Zhang, Yuan Yan Tang, Bin Fang, Yong Xiang, Document clustering in correlation similarity measure space, IEEE Trans. Knowl. Data Eng. 24 (6) (2012) 1002–1013.

[56] Wen Zhang, Taketoshi Yoshida, Xijin Tang, Qing Wang, Text clustering using frequent itemsets, Knowl.-Based Syst. 23 (5) (2010) 379–388.

[57] X. Zhang, L. Jing, X. Hu, M. Ng, X. Zhou, A comparative study of ontology based term similarity measures on pubmed document clustering, Advances in Databases: Concepts, Systems and Applications, 2010, pp. 115–126.

[58] Y. Zhao, G. Karypis, Criterion functions for document clustering: experiments and analysis, Technical Report 01-40, University of Minnestoa, 2001.

[59] Ying Zhao, George Karypis, Usama Fayyad, Hierarchical clustering algorithms for document datasets, Data Min. Knowl. Discov. 10 (2005) 141–168, http://dx.doi.org/10.1007/s10618-005-0361-3.