Clustering and Understanding Documents via Discrimination Information Maximization

Malik Tahir Hassan and Asim Karim

Dept. of Computer Science, LUMS School of Science and Engineering
Lahore, Pakistan
{mhassan,akarim}@lums.edu.pk

Abstract. Text document clustering is a popular task for understanding and summarizing large document collections. Besides the need for efficiency, document clustering methods should produce clusters that are readily understandable as collections of documents relating to particular contexts or topics. Existing clustering methods often ignore term-document semantics while relying upon geometric similarity measures. In this paper, we present an efficient iterative partitional clustering method, CDIM, that maximizes the sum of discrimination information provided by documents. The discrimination information of a document is computed from the discrimination information provided by the terms in it, and term discrimination information is estimated from the currently labeled document collection. A key advantage of CDIM is that its clusters are describable by their highly discriminating terms – terms with high semantic relatedness to their clusters' contexts. We evaluate CDIM both qualitatively and quantitatively on ten text data sets. In clustering quality evaluation, we find that CDIM produces high-quality clusters superior to those generated by the best methods. We also demonstrate the understandability provided by CDIM, suggesting its suitability for practical document clustering.

1 Introduction

Text document clustering discovers groups of related documents in large document collections. It achieves this by optimizing an objective function defined over the entire data collection. The importance of document clustering 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 in better document organization (e.g. digital libraries, corporate documents, etc) 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 readily understandable clusters. These requirements are often not satisfied in popular clustering methods. For example, in K-means clustering, documents are compared in the term space, which is 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.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 566–577, 2012.

In this paper, we present a new document clustering method based on discrimination information maximization (CDIM). CDIM's semantically motivated objective function is maximized via an efficient iterative procedure that 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 interpreted by a list of highly discriminating terms. These terms serve as units of understanding, as demonstrated in linguistics studies [1,2], describing a cluster in the document collection. We evaluate the performance of CDIM on ten popular text data sets. In clustering quality evaluation, CDIM is found to produce high quality clusters superior to those produced by non-negative matrix factorization (NMF) and several K-means variants. 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 work and motivation for our method in Section 2. CDIM, our document clustering method is described in detail in Section 3. Section 4 presents our experimental setup. Section 5 discusses the results of our experiments, and we conclude with future directions in Section 6.

2 Motivation and Related Work

Content-based document clustering continues to be challenging 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 [3]. It is a partitional clustering method that optimizes an objective function via an iterative two-step procedure. Usually, documents are represented by terms' weights, and documents are compared in the term space by the cosine similarity measure. Several clustering objective functions can be optimized [4] with the traditional objective of maximizing the similarity of documents to their cluster means producing reliable clusterings. 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) [5]. 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.

Some researchers have used external knowledge bases to semantically enrich the document representation for document clustering [6,7]. In [6], Wikipedia's concepts and categories are adopted to enhance the document representation, while in [7] several ontology-based (e.g. WordNet) term relatedness measures are evaluated for semantically smoothing the document representation. In both works, it has been shown that

the quality of clusterings produced by the K-means algorithm improves over the baseline ("bag of words") document representation. However, extracting information from knowledge bases is computationally expensive. Furthermore, 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 been tackled by clustering in a lower dimensional space of the original term space. One way to achieve this is through Non-Negative Matrix Factorization (NMF). NMF approximates the term-document matrix by the product of term-cluster and document-cluster matrices [8]. Extensions to this idea, with the goal of improving the interpretability of the extracted clusters, have also been proposed [9,10]. Another way is to combine clustering with dimensionality reduction techniques [11,12]. Nonetheless, these methods are restricted by their focus on approximation rather than semantically useful clusters, and furthermore, dimensionality reduction based techniques are often computationally expensive.

Recently, it has been demonstrated that the relatedness of a term to a context or topic in a document collection can be quantified by its discrimination information [2]. Such a notion of relatedness, as opposed to the traditional term-to-term relatedness, can be effectively used for data mining tasks like classification [13]. Meanwhile, measures of discrimination information, such as relative risk, odds ratio, risk difference, and Kullback-Leibler divergence, are gaining popularity in data mining [14,15]. In the biomedical domain, on the other hand, measures like relative risk have been used for a long time for cohort studies and factor analysis [16,17].

3 CDIM – Our Document Clustering Method

CDIM (Clustering via Discrimination Information Maximization) is an iterative partitional document clustering method that finds K groups of documents in a K-dimensional discrimination information 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 the ith document $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{Mi}]^T$ is represented by an M-dimensional vector (ith column of matrix $\mathbf{X}$). M is the total number of distinct terms in the N documents. The weight of term j in document i, denoted by x_{ji}, is equal to the count of term j in document i.

Our goal is to find K (usually in practice $K \ll \min\{M, N\}$) clusters $\mathcal{C}_k$ ($k = 1, 2, \ldots, K$) of documents such that if a document $\mathbf{x} \in \mathcal{C}_k$ then $\mathbf{x} \notin \mathcal{C}_j, \forall j \neq k$. 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 $\mathcal{T}_k$ 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 i for cluster k by d_{ik} and the discrimination information provided by document i for all clusters but cluster k by $\bar{d}_{ik}$, then the discrimination score of document i for cluster k is defined as $\hat{d}_{ik} = d_{ik} - \bar{d}_{ik}$. CDIM's objective function can then be written as

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

where $r_{ik} = 1$ if document 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 following 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 $r_{ik}, \forall i, k$, 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. 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 guaranteed because J is non-decreasing from one iteration to the next and J is upper-bounded by a local maxima.

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 for cluster k is quantified with the relative risk of the term for cluster k over the remaining clusters. Mathematically, the discrimination information of term j for cluster k and term j for all clusters but k is given by

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

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

where $p(x_j|\mathcal{C}_k)$ is the conditional probability of term j in cluster k and $\bar{\mathcal{C}}_k$ denotes all clusters but cluster k. The term discrimination information is either zero (no discrimination information) or greater than one with a larger value signifying higher discriminative power. The conditional probabilities in Equations 2 and 3 are estimated via smoothed maximum likelihood estimation.

3.4 Relatedness of Terms to Clusters

In Equations 2 and 3, $t \geq 0$ is a term selection parameter that controls the exclusion of terms that provide insignificant discrimination information. As the value of t is increased from zero, fewer terms will have a discrimination information greater than one.

The index set of terms that provide significant discrimination information for cluster k ($\mathcal{T}_k$) is defined as $\mathcal{T}_k = \{j|w_{jk} > 0, \forall j\}$. 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, $\mathcal{T}_k \cap \mathcal{T}_j \neq \emptyset, \forall j \neq k$. 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 [1], 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 [2]. Given these studies, we can consider all terms in $\mathcal{T}_k$ 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.

3.5 Document Discrimination Information

A document i is describable by the terms it contains. Each term 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 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 i for cluster k can be computed as the average term discrimination information for cluster k:

$$d_{ik} = \frac{\sum_{j \in \mathcal{T}_k} x_{ji} w_{jk}}{\sum_{j \in \mathcal{T}_k} x_{ji}}. \tag{4}$$

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 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 i belongs to cluster k. Note that a term contributes to the discrimination information of document i for cluster k only if it belongs to $\mathcal{T}_k$ and it occurs in document 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.

3.6 Algorithm

CDIM can be described 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 $w_{jk}, \forall j, k$ ($\bar{w}_{jk}, \forall j, k$), $\hat{\mathbf{D}}$ 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 $r_{ik}, \forall i, k$. 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

$$\hat{\mathbf{D}} = (\mathbf{X}\boldsymbol{\Sigma})^T(\mathbf{W} - \bar{\mathbf{W}}) \tag{5}$$

where $\boldsymbol{\Sigma}$ 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 i is assigned to cluster k if $\hat{d}_{ik} \geq \hat{d}_{ij}, \forall j \neq k$ (ties are broken arbitrarily). In matrix notation, this operation can be written as

$$\mathbf{R} = \text{maxrow}(\hat{\mathbf{D}}) \tag{6}$$

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 Equations 5 and 6 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 $\mathbf{W}$. It is easy to see that 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.

4 Experimental Setup

Our evaluations comprise of two sets of experiments. First, we evaluate the clustering quality of CDIM and compare it with other clustering methods on 10 text data sets. Second, we illustrate the understanding that is provided by CDIM clustering. The results of these experiments are given in the next section. Here, we describe our experimental setup.

Algorithm 1. CDIM – Document Clustering via Discrimination Information Maximization

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

4.1 Data Sets

Our experiments are conducted on 10 standard text data sets of different sizes, contexts, and complexities. The key characteristics of these data sets are given in Table 1. 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 10 are obtained from Karypis Lab, University of Minnesota[3]. Data sets 1 (stopword removal) and 3 to 10 (stopword removal and stemming) are available in preprocessed formats, while we perform stopword removal and stemming of data set 2. For more details on these standard data sets, please refer to the links given above.

4.2 Comparison Methods

We compare CDIM with five clustering methods. Four of them are K-means variants and one of them is based on Non-Negative Matrix Factorization (NMF) [8].

The four K-means variants are selected from the CLUTO Toolkit [18] based on their strong performances reported in the literature [5,3]. Two of them are direct K-way clustering methods while the remaining two are repeated bisection methods. For

[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

Table 1. Data sets and their characteristics

#	Name	Documents (N)	Terms (M)	Categories (K)
1	pu	672	19868	2
2	movie	1200	38408	2
3	reviews	4069	23220	5
4	hitech	2301	13170	6
5	tr31	927	10128	7
6	tr41	878	7454	10
7	ohscal	11162	11465	10
8	re0	1504	2886	13
9	wap	1560	8460	20
10	re1	1657	3758	25

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[4]. Specifically, we use the multiplicative update rule with Euclidean measure for approximating the term-document matrix.

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. For NMF, the term-document matrix is defined by term frequency values.

4.3 Clustering Validation Measures

We evaluate clustering quality with the BCubed metric [19]. In [20], it has been shown that the BCubed precision and recall are the only measures that satisfy all desirable constraints for a good clustering validation measure.

The BCUbed F-measure is computed as follows. 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 is equal to one, $Correct(o, o') = 1$, iff $L(o) = L(o') \leftrightarrow C(o) = C(o')$; otherwise $Correct(o, o') = 0$. BCubed precision (BP) and BCubed recall (BR) can now be defined as: $BP = Avg_o[Avg_{o'.C(o)=C(o')}[Correct(o, o')]]$ and $BR = Avg_o[Avg_{o'.L(o)=L(o')}[Correct(o, o')]]$. The BCubed F-measure is then given by $BF = 2 \times \frac{BP \times BR}{BP+BR}$. The BCubed F-measure ($BF$) ranges from 0 to 1 with larger values signifying better clusterings.

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

5 Results and Discussion

5.1 Clustering Quality

Table 2 gives the results of the clustering quality evaluation. 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). The shown values are average BCubed F-measure $\pm$ standard deviation, computed from 10 generated clusterings starting with random initial partitions.

These results show that CDIM outperforms the other algorithms on the ten data sets with five highest performance scores (shown in bold) and within 0.005 of the highest scores on three more data sets. CDIM is much better than NMF while its performances are closer to those of the K-means variants. We verified the consistency of these results using the Freidman's test, which is a non-parametric test recommended for evaluating multiple algorithms on multiple data sets [21]. At 0.05 significance level, CDIM is found to be significantly better than Direct-H2, RB-H2, and NMF, while its performance difference with Direct-I2 and RB-I2 is not statistically significant at this level.

An observation from our analysis is that CDIM consistently produces higher quality clusterings when the desired number of clusters is small (e.g. $K < 5$). 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 for document assignment. One potential way to overcome this shortcoming for larger number of clusters is to use a repeated bisection approach rather than a direct K-way partitioning approach.

Table 2. Clustering quality evaluation (average BCubed F-measure $\pm$ standard deviation)

Data	CDIM	Direct-I2	Direct-H2	RB-I2	RB-H2	NMF
pu	**0.706±0.06**	0.565±0.02	0.553±0.02	0.565±0.02	0.553±0.02	0.612±0.04
movie	**0.581±0.02**	0.533±0.02	0.522±0.01	0.533±0.02	0.522±0.01	0.510±0.01
reviews	0.667±0.05	0.627±0.06	0.626±0.06	0.609±0.04	**0.669±0.03**	0.552±0.03
hitech	**0.433±0.04**	0.391±0.02	0.380±0.02	0.394±0.02	0.390±0.03	0.399±0.02
tr31	**0.636±0.11**	0.585±0.05	0.575±0.05	0.553±0.07	0.572±0.05	0.362±0.03
tr41	0.603±0.05	**0.608±0.02**	0.584±0.03	0.602±0.05	0.590±0.04	0.361±0.04
ohscal	0.429±0.02	0.422±0.02	0.417±0.03	**0.432±0.01**	0.427±0.01	0.250±0.02
re0	**0.417±0.02**	0.382±0.02	0.382±0.01	0.397±0.03	0.375±0.01	0.345±0.02
wap	0.442±0.05	0.462±0.01	0.444±0.01	**0.465±0.02**	0.438±0.02	0.299±0.02
re1	0.393±0.03	**0.443±0.02**	0.436±0.02	0.416±0.01	0.418±0.03	0.301±0.03

5.2 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. We illustrate the understanding provided by CDIM's output by displaying

Table 3. Top 10 most discriminating terms (stemmed words) for clusters in ohscal data set

k	**Top 10 terms in cluster k**
1	'platelet', 'kg', 'mg', 'dose', 'min', 'plasma', 'pressur', 'flow', 'microgram', 'antagonist'
2	'carcinoma', 'tumor', 'cancer', 'surviv', 'chemotherapi', 'stage', 'recurr', 'malign', 'resect', 'therapi'
3	'antibodi', 'antigen', 'viru', 'anti', 'infect', 'hiv', 'monoclon', 'ig', 'immun', 'sera'
4	'patient', 'complic', 'surgeri', 'ventricular', 'infarct', 'oper', 'eye', 'coronari', 'cardiac', 'morta'
5	'pregnanc', 'fetal', 'gestat', 'matern', 'women', 'infant', 'deliveri', 'birth', 'labor', 'pregnant'
6	'risk', 'alcohol', 'age', 'children', 'cholesterol', 'health', 'factor', 'women', 'preval', 'popul'
7	'gene', 'sequenc', 'dna', 'mutat', 'protein', 'chromosom', 'transcript', 'rna', 'amino', 'structur'
8	'contract', 'muscle', 'relax', 'microm', 'calcium', 'effect', 'respons', 'antagonist', 'releas', 'action'
9	'il', 'receptor', 'cell', 'stimul', 'bind', 'growth', 'gamma', 'alpha', 'insulin', '0'
10	'ct', 'imag', 'comput', 'tomographi', 'scan', 'lesion', 'magnet', 'reson', 'cerebr', 'tomograph'

the top 10 most discriminating terms (stemmed words) for each cluster of the ohscal data set in Table 3. 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). By looking at the top ten terms, it is easy to determine the category of most clusters: 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 2 and 8 still give coherent meaning to the documents they contain.

As another example, in hitech data set, the top 5 terms for two clusters are: (1) 'health', 'care', 'patient', 'hospit', 'medic', and (2) 'citi', 'council', 'project', 'build', 'water'. The first cluster can be mapped to the health category while the second cluster does not have an unambiguous mapping to a category but it still gives sufficient indication that these articles discuss hi-tech related development projects.

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 Figure 1 (left plot). The x- and y-axes in this plot correspond to document discrimination information for cluster 1 and 2 (d_{i1} and d_{i2}), respectively. and the colored makers give the true categories. It is seen that the two clusters are spread along the two axes and the vast majority of documents in each cluster belong to the same category. Similar scatter plots for Direct-I2 and NMF are shown in the middle and right plots, respectively, of Figure 1. However, these methods exhibit poor separation between the two categories in the pu data set.

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.

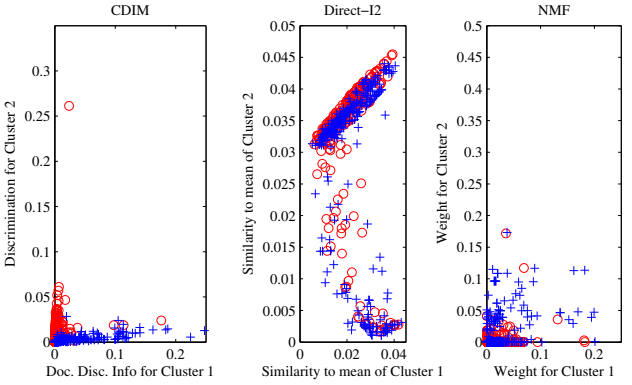

Fig. 1. Scatter plot of documents projected onto the 2-D discrimination information space (CDIM), similarity to cluster mean space (Direct-I2), and weight space (NMF). True labels are indicated by different color markers.

6 Conclusion and Future Work

In this paper, we propose and evaluate a new document clustering method, CDIM, that 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 respective 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 is estimated from the document collection via its relative risk. 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.

Our experimental evaluations confirm the effectiveness of CDIM as 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 particular, we would like to investigate other measures of discrimination/relatedness information, extend and evaluate CDIM for soft clustering, and develop a hierarchical and repeated bisection version of CDIM.

References

1. Morris, J., Hirst, G.: Non-classical lexical semantic relations. In: Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, pp. 46–51. Association for Computational Linguistics (2004)
2. Cai, D., van Rijsbergen, C.J.: Learning semantic relatedness from term discrimination information. Expert Systems with Applications 36, 1860–1875 (2009)
3. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley, New York (2006)

4. Zhao, Y., Karypis, G.: Criterion functions for document clustering: experiments and analysis. Technical Report 01-40, University of Minnestoa (2001)
5. Steinbach, M., Karypis, G.: A comparison of document clustering techniques. In: Proceedings of the KDD Workshop on Text Mining (2000)
6. Hu, X., Zhang, X., Lu, C., Park, E., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 389–396. ACM (2009)
7. Zhang, X., Jing, L., Hu, X., Ng, M., Zhou, X.: A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 115–126. Springer, Heidelberg (2007)
8. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 267–273. ACM (2003)
9. Xu, W., Gong, Y.: Document clustering by concept factoriz ation. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 202–209. ACM (2004)
10. Cai, D., He, X., Han, J.: Locally consistent concept factorization for document clustering. IEEE Transactions on Knowledge and Data Engineering (2010)
11. Tang, B., Shepherd, M., Heywood, M.I., Luo, X.: Comparing Dimension Reduction Techniques for Document Clustering. In: Kégl, B., Lee, H.-H. (eds.) Canadian AI 2005. LNCS (LNAI), vol. 3501, pp. 292–296. Springer, Heidelberg (2005)
12. Ding, C., Li, T.: Adaptive dimension reduction using discriminant analysis and k-means clustering. In: Proceedings of the 24th International Conference on Machine Learning, pp. 521–528. ACM (2007)
13. Junejo, K., Karim, A.: A robust discriminative term weighting based linear discriminant method for text classification. In: Eighth IEEE International Conference on Data Mining, pp. 323–332 (2008)
14. Li, H., Li, J., Wong, L., Feng, M., Tan, Y.P.: Relative risk and odds ratio: a data mining perspective. In: PODS 2005: Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (2005)
15. Li, J., Liu, G., Wong, L.: Mining statistically important equivalence classes and delta-discriminative emerging patterns. In: KDD 2007: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2007)
16. Hsieh, D.A., Manski, C.F., McFadden, D.: Estimation of response probabilities from augmented retrospective observations. Journal of the American Statistical Association 80(391), 651–662 (1985)
17. LeBlanc, M., Crowley, J.: Relative risk trees for censored survival data. Biometrics 48(2), 411–425 (1992)
18. Karypis, G.: CLUTO-a clustering toolkit. Technical report, Dept. of Computer Science, University of Minnesota, Minneapolis (2002)
19. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, pp. 79–85. ACL (1998)
20. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval 12(4), 461–486 (2009)
21. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)