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/

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