Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 49 (2015) 118 – 127

Procedia
Computer Science

ICAC3'15

Discovering Context of Labeled Text Documents using Context Similarity Coefficient

Anagha Kulkarni[a,*], Vrinda Tokekar[b], Parag Kulkarni[c]

[a]Cummins College of Engineering for Women, Karvenagar, Pune 411052, India
[b]IET, DAVV, Khandwa Road, Indore 452017, India
[c]EkLat Research, Pune, India

Abstract

To find closeness between two data points, traditional distance based closeness measurement calculates distance between two data points. However, it fails to capture behaviour of data series. Behaviour of data series can be captured by association and disassociation between patterns of data points. This can reflect closeness between them. The same concept can be applied to find association between text documents. Using this philosophy, this paper proposes a novel approach of document association based on context similarity coefficient (CSC). CSC based document association helps to capture contextual relationship between documents. Experiments conducted on standard datasets such as Reuters-21578 and RCV1 show that CSC successfully finds closeness between the documents.

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)

Keywords: Context discovery; context closeness value; text mining; pattern based clustering

1. Introduction

Clustering is the task of building groups or clusters of similar objects. Most of the clustering algorithms find similarity between objects using distance. Clustering algorithms use distance functions such as Euclidean distance, Manhattan distance, Minkowski distance, cosine similarity etc. to group objects in clusters. The clusters are formed in such a way that distance between two objects within a cluster is minimum and that between different clusters is maximum. Clustering using distance functions, called distance based clustering, is a very popular technique to cluster the objects and has given good results. However, this type of clustering may not be suitable for all applications. Pattern based clustering is another way of clustering[1,2,3]. Pattern based clustering builds clusters based on similarity of patterns. Pattern based clustering is very important in applications such as trend analysis, climatology, gene expression, chromosome matching[2,4]. Music also has rhythmic or melodic patterns with different scales. There may exist correlations between two objects which are far away from each other distance wise but have similar patterns.

* Corresponding author. Tel.: +91-020-2531-1000 ; fax: +0-000-000-0000.
E-mail address: anagha.kulkarni@cumminscollege.in

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)
doi:10.1016/j.procs.2015.04.235

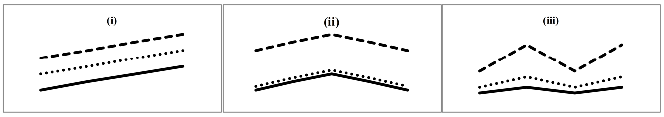

Fig. 1. Simple patterns

Sample patterns are shown in figure 1. In the first case, all the three objects are close to each other distance wise and have similar pattern. In second case, bottom two objects are very close to each other and all objects have similar pattern. The third object is shifted away. In the third case, third object is shifted. Even though, the patterns are similar, the objects are also scaled. The goal of this paper is to group objects in all such cases having similar pattern.

Human beings are the best pattern recognition machines. When articles, news, research papers and other text documents are written, the documents are bound to have patterns. With the increasing amount of unstructured text documents, it is very important to discover patterns within them. As far as text documents are concerned, one can find patterns in usage of words or terms (terms are words that are preprocessed: thus a document is a set of terms). Preprocessing of words primarily includes stop word removal and stemming. Many researchers have found patterns by mining frequent termsets (similar to itemsets)[5,6]. This paper uses a method based on closeness factor to find similar patterns[7].

For finding patterns from documents using terms, the documents are required to be represented using Vector Space Model (VSM). In VSM, every document is transformed into a vector in high-dimensional term space. Every distinct term in these documents represents a dimension. When the number of dimensions increase, the volume of space increases and the data becomes sparse. It becomes very difficult to find patterns in such cases. In addition, identified patterns may be incorrect. Thus, it is necessary to identify patterns using only relevant terms or dimensions. Many techniques have been applied to find relevant dimensions. Some of the modern techniques are sparse principal component analysis (SPCA)[8], regularized latent semantic indexing (RLSI)[9], imprecise spectrum analysis (ISA)[10] and so on.

Terms in a document give meaning or context to a document. In other words, they establish context of the document. Many researchers have defined 'context' based on the application. It is an *unsaid* or *unspecified* information. Author, readability index, language and so on form context of the document[11]. However, in the present paper, context is defined as follows:

Context is a secondary source of information which is not directly specified. It may be the data that comes from outside. From document's perspective, it could be other terms that surround a given term, the positions of the terms in the document and the metadata of the document.

In this paper, WordNet is used to find relevant terms. Using terms from WordNet, every document is transformed into a vector in low-dimensional term space. This is called context vector. The most important advantage of using low-dimensional term space over high-dimensional term space is that it is computationally efficient. At the same time, the patterns are clear. Using context vector, context similarity coefficient (CSC) between the documents is found. CSC is used to group documents to form clusters. A very simple technique for clustering is devised which is based on threshold. Documents having similar context vector pattern are grouped in a cluster. The most important advantage of this clustering technique is that it works without a priori knowledge of number of clusters.

The paper is organized as follows. Section 2 discusses CSC. Section 3 demonstrates how documents are clustered using CSC. Section 4 discusses and analyzes results. Detailed discussion of proposed methodologies is presented in 5. Conclusions are stated in section 6.

2. Context Similarity Coefficient (CSC)

As mentioned earlier, context is an *unsaid* or *unspecified* information. Metadata of a document plays a very important role in deciding the context of the document. Author of the document, language, date on which the document is written etc are popularly considered as metadata of a document. In text mining, documents are assigned categories after text categorization. However, the documents having same category, may have different contexts. The context of