Dynamic Clustering Based on Minimum Spanning Tree and Context Similarity for Enhancing Document Classification

Anirban Chakrabarty, Department of Computer Applications, Future Institute of Engineering and Management, Sonarpur, Kolkata, India

Sudipta Roy, Triguna Sen School of Information Technology, Assam Central University, Silchar, Assam, India

ABSTRACT

Document Classification is the task of assigning a text document to one or more predefined categories according to its content and the labeled training samples. Traditional classification schemes use all training samples for classification, thereby increasing storage requirements and calculation complexity as the number of features increase. Moreover, the commonly used classification techniques consider the number of categories is known in advance, this may not be so in actual reality. In the practical scenario, it is very much essential to find the number of clusters for unknown dataset dynamically. Identifying these limitations, the proposed work evolves a text clustering algorithm where clusters are generated dynamically based on minimum spanning tree incorporating semantic features. The proposed model can efficiently find the significant matching concepts between documents and can perform multi category classification. The formal analysis is supported by applications to email and cancer data sets. The cluster quality and accuracy values were compared with some of the widely used text clustering techniques which showed the efficiency of the proposed approach.

Keywords: Minimum Spanning Tree, Spam Filtering, Dynamic Clustering, Context Similarity

INTRODUCTION

In the digital era, the rapid growth in the volume of text documents available from various sources like Internet, digital libraries, medical records has spurred users to effectively retrieve, navigate, and organize information. The ultimate goal is to help users to search what they are looking for effortlessly and take decisions suitably. In this context, fast and high-quality document clustering algorithms play a major role. Most of the common techniques in text retrieval are based on the statistical analysis of terms i.e. words or phrases. Such text retrieval methods are based on the vector space model (VSM) which is a widely used data representa-
tion. The VSM represents each document as a feature vector of the terms in the form of term frequency or term weight (Salton et al., 1975). The similarity between documents is measured by one of the several similarity measures that are based on feature vector. Examples include the cosine measure and the Jaccard measure (Schaeffer, 2007). Metric distances such as Euclidean distance are not appropriate for high dimension and sparse domains. Most conventional measures estimate the surface overlap between documents based on the words they mention and ignore deeper semantic connections. To achieve a more accurate analysis, the underlying model should indicate the semantics of text. Conceptual information retrieval extracts information by processing the document on semantic level forming a concept base and then retrieves relative information to provide search results.

Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. Methods used for text clustering include decision trees, contextual clustering, clustering based on data summarization, statistical analysis, neural nets and rule-based systems among others (Nahm & Mooney, 2000; L. Talavera & J. Bejar, 2001; Jin et al., 2005). Most clustering techniques consider the number of clusters is fixed which can result in poor quality clustering. Dynamic document clustering is the process of inserting the newly arrived documents to the appropriate existing cluster so it is not required to relocate clusters, thus time and effort taken for clustering is drastically reduced (Wang et al., 2011; Nadig et al., 2008). Figure 1 shown below depicts a model for dynamic document clustering.

A spanning tree is an acyclic sub graph of a graph G, which contains all vertices from G and is also a tree. The minimum spanning tree (MST) of a weighted graph is the minimum weight spanning tree of that graph (Edla & Jana, 2013). MST clustering algorithm is known to be capable of detecting clusters with irregular boundaries. Moreover MST is relatively insensitive to small amounts of noise spread over the field (Zahn, 1971). Thus the shape of a cluster boundary has little impact on the performance of the algorithm. The proposed approach does not require a preset number of clusters. Edges that satisfy a predefined inconsistency measure are removed from the tree. The process is iterated until there is a change in the edge list and all data are clustered.

The paper suggests a context based retrieval method at the sentence, document and corpus levels for enhancing the quality of text retrieval. More specifically, it can quantify how closely concepts relate to each other and integrate this into a document similarity measure. As a result, documents do not have to mention the same words to be judged similar. The suggested clustering technique is applied on two different data sets for developing clusters - email messages and cancer data sets to demonstrate its feasibility. A major contribution of this work is in developing clusters dynamically based on area of interest of email users and when applied to cancer data sets it can classify patients to different treatment clusters based on age groups. The work introduces a text classification algorithm which allows incremental and multi label classification by comparing with the context pool i.e. the most significant concepts in the cluster.

The subsequent sections of the paper are organized as follows: A brief literature review on related classification approaches is presented in Section-2. The problem definition follows in Section-3. The methodology and aim of the research work is depicted in Section-4. Section–5 presents some important concepts associated with the proposed model. Section-6 proposes an algorithm for developing the clusters using a minimum spanning tree and performs time complexity analysis. In later sections evaluation parameters are developed to compare the clustering results with other established state of art methods.

Related Work

A tree is a widely used data structure in the field of information retrieval and studies on tree similarity algorithms are not new. Computation of tree edit distance is focused on finding
Conclusions and Further Research
Ibrahim Dweib and Joan Lu (2013). Design, Performance, and Analysis of Innovative Information Retrieval (pp. 264-269).
www.igi-global.com/chapter/conclusions-further-research/69142?camid=4v1a