A Dynamic and Semantically-Aware Technique for Document Clustering in Biomedical Literature

Min Song, New Jersey Institute of Technology, USA
Xiaohua Hu, Drexel University, USA
Illhoi Yoo, University of Missouri, USA
Eric Koppel, New Jersey Institute of Technology, USA

ABSTRACT

As an unsupervised learning process, document clustering has been used to improve information retrieval performance by grouping similar documents and to help text mining approaches by providing a high-quality input for them. In this article, the authors propose a novel hybrid clustering technique that incorporates semantic smoothing of document models into a neural network framework. Recently, it has been reported that the semantic smoothing model enhances the retrieval quality in Information Retrieval (IR). Inspired by that, the authors developed and applied a context-sensitive semantic smoothing model to boost accuracy of clustering that is generated by a dynamic growing cell structure algorithm, a variation of the neural network technique. They evaluated the proposed technique on biomedical article sets from MEDLINE, the largest biomedical digital library in the world. Their experimental evaluations show that the proposed algorithm significantly improves the clustering quality over the traditional clustering techniques including k-means and self-organizing map (SOM).

Keywords: document clustering, feature selection, neural network

INTRODUCTION

Document clustering, unlike document classification, is an unsupervised learning process meaning that there is no known information about documents including the number of document groups (usually called k). Document clustering organizes textual documents into meaningful groups that represent topics in document collections without any known information about a document set. As a result, the documents in a document cluster are similar

DOI: 10.4018/jdwm.2009080703
to one another while documents from different clusters are dissimilar.

Document clustering was originally studied to enhance the performance of information retrieval (IR) because similar documents tend to be relevant to the same user queries (Wang et al., 2002; Zamir & Etzioni, 1998). Document clustering has been used to facilitate nearest-neighbor search (Buckley & Lewit, 1985), to support an interactive document browsing paradigm (Cutting et al., 1992; Gruber, 1993; Koller & Sahami, 1997; Gruber, 1993), and to construct hierarchical topic structures (van Rijsbergen, 1979). Thus, document clustering plays a more important role for IR and text mining communities since the most natural form for storing information is text, and text information has increased exponentially.

In the biomedical domain, document clustering technologies have been used to facilitate the practice of evidence-based medicine. This is because document clustering enhances biomedical literature searching (e.g., MEDLINE searching) in several ways and literature searches are one of the core skills required for the practice of evidence-based medicine (Evidence-based Medicine Working Group, 1992). For example, Pratt and her colleagues (Pratt et al., 1999; Pratt & Fagan, 2000), and Lin and Demner-Fushman (2007) introduced interesting semantic document clustering approaches that automatically cluster biomedical literature (MEDLINE) search results into document groups for the better understanding of literature search results.

Current information technologies allow us to acquire, store, archive, and retrieve documents electronically. To this end, document clustering has been given focal attention because document clustering assists users in discovering hidden similarities and key concepts in documents. One of most serious problems making document clustering difficult to deal with text information is that the size of text collections in digital libraries are increasing rapidly. To handle the increasing size of document collections, a clustering algorithm has to not only solve the incremental problem but it must also have high efficiency in a large dataset.

Most document clustering algorithms require a form of data pre-processing including stop-word removal and feature selection. Through the data pre-processing, unimportant features are eliminated and the original dimension is reduced to a more manageable size. However, the data pre-processing has two problems. First, although the data pre-processing can reduce the original dimension size, the reduced dimension is still sparse, which is called “the curse of dimensionality”. As the result, clustering results are often low quality. Second, the reduction of dimensionality by the data-preprocessing may disturb the preservation of the original topological structure of the input data.

To solve these problems, we propose a context-sensitive semantic smoothing of a document model and incorporate it into Dynamic Growing Cell Structure (DynGCS). The effect of model smoothing has not been extensively studied in the context of document clustering (Zhang et al., 2006). Most model-based clustering approaches simply use Laplacian smoothing to prevent zero probability (Nigam & McCallum, 1998; Zhong & Ghosh, 2005), while most similarity-based clustering approaches employ the heuristic TF*IDF scheme to discount the effect of general words (Steinbach et al., 2000). As showed in (Zhong & Ghosh, 2005), model-based clustering has several advantages over discriminative based approaches. One of the advantages of model-based approaches is that it learns generative models from the documents, with each model representing one particular document set. Due to the promising results reported in model-based clustering approaches, we propose a novel semantic smoothing technique to improve clustering quality.

DynGCS is an adaptive variant of an artificial neural network model, Self-Organizing Map (SOM), which is well suited for mapping high-dimensional data into a 2-dimensional representation space. The training process is based on weight vector adaptation with respect to the input vectors. SOM has shown to be a highly
Related Content

Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse
www.irma-international.org/article/finding-non-coincidental-sporadic-rules/1765/

Future Research Directions in E-Tourism Studies: Blind Spots and Complaint Analyses Using Data Science Method
www.irma-international.org/chapter/future-research-directions-in-e-tourism-studies/150269/

Cluster-Based Input Selection for Transparent Fuzzy Modeling
www.irma-international.org/article/cluster-based-input-selection-transparent/1771/

Preference-Based Frequent Pattern Mining
www.irma-international.org/article/preference-based-frequent-pattern-mining/1759/

Dynamic View Selection for OLAP
www.irma-international.org/chapter/dynamic-view-selection-olap/40399/