Text Clustering based on Semantic Measure

by

Walaa Khaled Ebn El-Waleed

A thesis
presented to Ain Shams University
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer and Information Sciences

Cairo, Egypt, 2010

Abstract

Text document clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters. Most text clustering techniques are based on the statistical analysis of a term (word/phrase) to represent text documents. The statistical analysis captures the importance of the term based on its frequency weight which is the number of occurrence of this term in the document. The frequency weight is often unsatisfactory for document representation. It ignores the relationships between terms, and considers them as independent features. Therefore, there is an intensive need for a model that captures the semantic of the text in a formal structure. The underlying model should indicate terms that capture the semantics of text, and integrates background knowledge into the process of clustering text documents.

A new semantic-based model that utilizes WordNet in text clustering is introduced. The semantic-based model (SSM) can effectively represent text document semantically. The poroposed model discriminates between non-important terms and important terms with respect to semantics. The proposed semantic-based model is based on single word semantic similarity analysis (WSSM) and phrase semantic similarity analysis (PSSM) as well. The model adds new weights to document terms reflecting the semantic similarity between documents. The SSM assigns higher weights to terms that are semantically close. In our model, each document is analyzed to extract terms considering stemming and pruning processes. We adopt the extended gloss overlap measure to get the semantic relatedness between terms pairs.

Text documents contain general words (class-independent) and core words (class-specific). The core words represent the individual classes topic, and the general words have similar distributions on different classes. The general words cause the difficulty of document clustering. The proposed WSSM discounts the effect of the general words and aggravates the effects of the core terms. Moreover, the phrase semantic similarity model (PSSM) utilizes phrases as a more explicit term for analysis. The PSSM generates new semantic weights for phrases as a richer source of information in the document. In addition, it captures the semantic similarities between documents that have semantically similar terms but unnecessary syntactically identical.

We propose a new Semantic Similarity Histogram based Incremental Document Clustering (SHC). The proposed SHC integrates the semantic relationships between documents to the incremental clustering. The main objective is to maintain high cluster cohesiveness, when a new document is added to the dataset. Insertion order is a big challenge in the incremental clustering algorithms. Different document insertion orders will lead to different results for clustering quality. The SHC solves this problem by removing bad documents that reduce the cluster cohesiveness, and reassigning them to other more appropriate cluster. It is a negotiation protocol between clusters to increase the coherency as high as possible.

Results show that semantic-based similarity model has a significant clustering improvement. The experiments demonstrate extensive comparison between traditional document representation model (VSM), the semantic-based document representation obtained by the semantic model and other semantic methods. Experimental results demonstrate the substantial enhancement of the quality using: (1) Phrase only, (2) word semantic similarity model (WSSM), (3) Phrase semantic similarity model (PSSM). Moreover, the Semantic Similarity Histogram based Incremental Document Clustering (SHC) has a better performance than traditional clustering algorithms in terms of Entropy, F-measure, and Purity clustering quality measures.

Acknowledgements

This thesis would not be possible without the support of many individuals, to whom I would like to express my gratitude. I will always be indebted to my supervisors, Prof. Essam Khalifa, Prof. Mohamed Hashem, and Prof. Mohamed Kamel, for their support, encouragement, guidance, and most importantly trust.

My deepest gratitude is to my supervisor, Prof Essam Khalifa. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. He taught me how to question thoughts and express ideas. His patience and support helped me overcome many crisis situations and finish this dissertation.

I am indebted to the generous help of my co-supervisor Professor Mohamed Kamel for his support and provision of this work. He is a source of inspiration for innovative ideas, and his kind support is well known to all his students and colleagues.

Prof Mohamed Hashem's insightful comments and constructive criticisms at different stages of my research were thought-provoking and they helped me focus my ideas. I am grateful to him for holding me to a high research standard and enforcing strict validations for each research result, and thus teaching me how to do research.

I would like also to thank all my colleagues in the PAMI research group at the University of Waterloo. They have been helpful in many situations and the knowledge we shared with each other was so valuable to the work presented in this thesis.

Most importantly, none of this work would have been possible without the love and patience of my family, my husband Tamer and my daughter Maryam. I would like to thank my parents for their support and encouragement throughout my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the abundance of text documents available through corporate document management systems and the World Wide Web, the dynamic partitioning of texts into previously unseen categories is a major topic for many applications such as information retrieval from databases and business intelligence solutions. In every one of these applications information is involved. However, with this continuous growth of awareness and the corresponding growth of information, it has become clear that we need to organize information in such a way that will make it easier for everyone to access various types of information.

Text Mining is an automated technique that aims to discover high level information in huge amount of textual data. Clustering text documents into categories is an important step in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems. A well known challenge in text clustering is handling of text data in complex semantics and linguistics. Although some methods [1, 2, 3, 4, 5] are effectively dealing with complex semantics, this area is still a remaining problem.

Most traditional text clustering methods are based on the vector space model (VSM) representation, which is based on terms frequencies weights. The VSM ignores the important information on the semantic relationships between documents terms, and deals with them as independent features. To overcome this problem, many methods integrate WordNet to text clustering as an ontology to enrich text representations. WordNet is a vocabulary and a lexical ontology that attempts to model the lexical knowledge of a native speaker of English into a taxonomic hierarchy. Entries (i.e., terms or concepts) are organized into synsets (i.e., lists of synonym terms or concepts), which in turn are organized into senses (i.e., different meanings of the same term or concept). There are also different categorizations corresponding to nouns, verbs, adverbs etc. Each entry is related to entries higher or lower in the hierarchy by different types of relationships.

Most of the previous methods represent document semantically based on enrichment strategies. The enrichment strategy is to append or replace document terms with their hypernym and synonym. Although this strategy is simple, it is not applicable on large text datasets and causes to information loss.

1.1 Motivations

Text clustering is an unsupervised learning method which groups text document into related clusters, and discovers hidden knowledge between clusters. Text clustering has been applied to many applications as indexing, information retrieval, browsing large document collections and mining text data on the Web.

Most current document clustering approaches are based on the vector-space model (VSM), also called bag of words model or word space. The dimensions of the vector space are constituted by the important terms (usually words) of the document collection. The respective term or word frequencies (TF) in a given document constitute the vector describing this document. In order to discount frequent terms with little discriminating power, each term can additionally be weighted based on its Inverse Document Frequency (IDF) in the document collection. Once the documents are mapped into the vector space, they can be clustered according to the distances between the vectors.

This simple VSM model cannot represent text semantics because terms are considered as independent features. However, many terms in text are semantically related. For example, (suggestion, advice) and (meat, beef), can have the same meaning in a document but they will be represented as two independent terms. This situation can significantly affect the similarity calculation between two documents, and therefore affect clustering results. Indeed, the VSM model ignores all important semantic relations of terms, such as synonymy, specialisation/generalisation and part/whole relationships in forming the text document representation. Therefore, there is a need for a document representation model that captures the semantics in text in a formal structure.

Recently, WordNet as an ontology has been applied to various clustering methods to improve their performance. Most existing WordNet-based clustering methods are based on document vectors enrichment. It means that the terms synonymys are added to documents to calculate frequencies weights without considering other semantic relationships between terms. These methods are based on single term analysis and they do not perform any phrase analysis. Although, phrases can be utilized as a more informative terms for analysis.

The real motivation behind the work in this thesis is to create a semantic model that is able to categorize text documents effectively, based on a semantic representation of the document data, and targeted towards achieving high degree of clustering quality. The main goal is to increase the importance of cluster- dependent core terms and to reduce the contribution of cluster-independent general words so that the similarity between two documents can be calculated more accurately in text clustering and good clustering results can be obtained.

In addition, the proposed model aims at grouping text documents such that documents in each group are closely related to each other (where all the documents in the group are related to the same topic), while the documents from different groups should not be related to each other (i.e. of different topics). The similarity between the documents in one category (intra-cluster similarity) is maximized , and the similarity between different categories (inter-cluster similarity) is minimized. Consequently the quality of categorization (clustering) should be maximized. Figure 1.1 illustrates this concept.

The next step is to perform incremental clustering of the documents

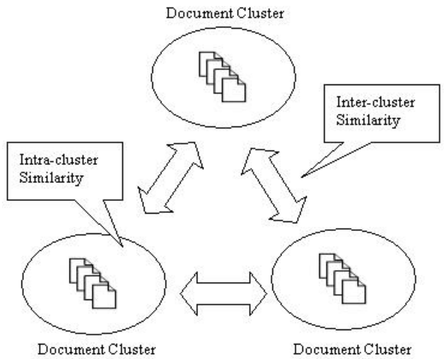

Figure 1.1: Intra-Cluster and Inter-Cluster Similarity

using a special cluster representation. The representation relies on a quality criteria called the Cluster Semantic Similarity Histogram that is introduced to represent clusters using the similarities between documents inside the clusters. Because the clustering technique is incremental, new documents being clustered are compared to cluster histograms, and are added to clusters such that the cluster similarity histograms are improved

1.2 Contributions

by classical document representation model, such as the vector space model, is based on plain lexicographic term matching between terms (eg. two documents are considered similar if they are lexicographically the same). However, plain lexicographic analysis and matching is not generally sufficient to determine if two terms are similar and consequently whether two documents are similar. Two terms can be lexicographically different but have the same meaning (eg. they are synonyms) or they may have approximately the same meaning (they are semantically similar).

Therefore, we propose a semantic similarity based model to represent documents with embedding of semantic relationship information of terms in the text documents representation. The proposed model embed the semantic relationship information directly in the weights of the corresponding terms which are semantically related by readjusting the