Chapter XI
Conceptual Clustering of Textual Documents and Some Insights for Knowledge Discovery

Leandro Krug Wives
Federal University of Rio Grande do Sul, Brazil

José Palazzo Moreira de Oliveira
Federal University of Rio Grande do Sul, Brazil

Stanley Loh
Catholic University of Pelotas, Brazil
Lutheran University of Brazil, Brazil

ABSTRACT

This chapter introduces a technique to cluster textual documents using concepts. Document clustering is a technique capable of organizing large amounts of documents in clusters of related information, which helps the localization of relevant information. Traditional document clustering techniques use words to represent the contents of the documents and the use of words may cause semantic mistakes. Concepts, instead, represent real world events and objects, and people employ them to express ideas, thoughts, opinions, and intentions. Thus, concepts are more appropriate to represent the contents of a document and its use helps the comprehension of large document collections, since it is possible to summarize each cluster and rapidly identify its contents (i.e., concepts). To perform this task, the chapter presents a methodology to cluster documents using concepts and presents some practical experiments in a case study to demonstrate that the proposed approach achieves better results than the use of words.
INTRODUCTION

Organizations are producing, collecting, and storing more and more documents, most of them containing texts. According to Tan (1999), about 80% of the documents an organization has are in textual format. This huge volume of unstructured or semi-structured data turns the act of finding relevant information harder to accomplish, generating a problem known as the information overload problem (Chen, 1996).

Techniques and tools from knowledge discovery in texts (KDT) (Feldman & Dagan, 1995) or simply text mining (Tan, 1999) are being developed to cope with this problem. One of these techniques is clustering (i.e., cluster analysis), a technique used to group similar documents of a given collection, helping the comprehension of its contents (Jain, Murty, & Flynn, 1999; Lu, Motoda, & Liu, 1997; Willet, 1988). One of its goals is to put similar documents in the same cluster and to place dissimilar documents in different clusters. The hypothesis is that, through a clustering process, similar objects will remain in the same group according to the attributes they have in common. This hypothesis is known as the cluster hypothesis, and it is appropriately described by Rijsbergen (1979).

It is important to analyze the difference between clustering and classification processes. The two activities have different purposes and are different in essence: the clustering process generates clusters without previous knowledge about document contents or about which classes exist, while classification (or categorization) is a process that starts with a predefined set of categories and tries to identify to which class a document belongs (Willet, 1988). Thus, clustering helps people identify classes and build taxonomies.

Some benefits of the clustering process are:

- Impartiality in the process: Humans have background knowledge that tends to bias the personal clustering process;
- Identification of common characteristics among documents (patterns such as words or concepts): They can be used to understand or explain why the documents, within a cluster, are similar and to understand a whole collection of documents and its contents;
- Organization of unstructured documents: Clustering minimizes the information overload (incapacity to analyze excessive amounts of information) in information retrieval, automatically grouping similar information (splitting a set of documents), summarizing its common characteristics, thus helping the user to visualize the results of a query or search, and aiding the user to navigate in a set of documents.

The clustering of textual documents, specifically Web pages, is one of the current research challenges. This task is comparable to the clustering of structured records in a database. In both cases, elements are compared by means of a similarity function, which analyzes the elements’ characteristics. However, when dealing with textual documents as elements, there is the need of an additional (previous) step to identify the characteristics of the documents, since these characteristics are not as clear as in structured data. Traditional textual clustering approaches use words or keywords, extracted automatically from the text, to represent the elements. However, words can lead to semantic mistakes, known as the vocabulary problem (Chen, 1994). It happens, for example, when people use subjective language facets, as synonymy (different words for the same meaning), polysemy (the same word with many meanings), lemmas (words with the same radical, like the verb “to marry” and the noun “marriage”) and quasisynonymy (semantically related words).
Related Content

Sequential Patterns Postprocessing for Structural Relation Patterns Mining
[www.igi-global.com/article/sequential-patterns-postprocessing-structural-relation/1814?camid=4v1a](www.igi-global.com/article/sequential-patterns-postprocessing-structural-relation/1814?camid=4v1a)

Weak Ratio Rules: A Generalized Boolean Association Rules
Baoqing Jiang, Xiaohua Hu, Qing Wei, Jingjing Song, Chong Han and Meng Liang (2011). *International Journal of Data Warehousing and Mining* (pp. 50-87).
[www.igi-global.com/article/weak-ratio-rules/55079?camid=4v1a](www.igi-global.com/article/weak-ratio-rules/55079?camid=4v1a)

Information Security Effectiveness: Conceptualization and Validation of a Theory
[www.igi-global.com/chapter/information-security-effectiveness/29146?camid=4v1a](www.igi-global.com/chapter/information-security-effectiveness/29146?camid=4v1a)

Data Field for Hierarchical Clustering
Shuliang Wang, Wenyan Gan, Deyi Li and Deren Li (2011). *International Journal of Data Warehousing and Mining* (pp. 43-63).
[www.igi-global.com/article/data-field-hierarchical-clustering/58637?camid=4v1a](www.igi-global.com/article/data-field-hierarchical-clustering/58637?camid=4v1a)