IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB8927

Chapter VIII

Mining Text Documents for Thematic Hierarchies Using Self-Organizing Maps

Hsin-Chang Yang
Chang Jung University, Taiwan

Chung-Hong Lee
Chang Jung University, Taiwan

ABSTRACT

Recently, many approaches have been devised for mining various kinds of knowledge from texts. One important application of text mining is to identify themes and the semantic relations among these themes for text categorization. Traditionally, these themes were arranged in a hierarchical manner to achieve effective searching and indexing as well as easy comprehension for human beings. The determination of category themes and their hierarchical structures was mostly done by human experts. In this work, we developed an approach to automatically generate category themes and reveal the hierarchical structure among them. We also used the generated structure to categorize text documents. The document collection was trained by a self-organizing map to form two feature maps. We then analyzed these maps and obtained the category themes and their structure. Although the test corpus contains documents written in Chinese, the proposed approach can be applied to documents written in any language, and such documents can be transformed into a list of separated terms.

This chapter appears in the book, *Data Mining: Opportunities and Challenges,* edited by John Wang.

INTRODUCTION

In text categorization, we try to assign a text document to some predefined category. When a set of documents is well categorized, both storage and retrieval of these documents can be effectively achieved. A primary characteristic of text categorization is that a category reveals the common theme of those documents under this category; that is, these documents form a natural cluster of similar context. Thus, text categorization provides some knowledge about the document collection. An interesting argument about text categorization is that before we can acquire knowledge through text categorization, we need some kinds of knowledge to correctly categorize documents. For example, two kinds of key knowledge we need to perform text categorization are 1) the categories that we can use, and 2) the relationships among the categories. The first kind of knowledge provides a set of themes that we can use to categorize documents. Similar documents will be categorized under the same category if they have the same theme. These categories form the basis of text categorization. The second kind of knowledge reveals the structure among categories according to their semantic similarities. Ideally, similar categories, i.e., categories with similar themes, will be arranged "closely" within the structure in some manner. Such arrangement provides us with an effective way to store and retrieve documents. Moreover, such structure may make the categorization result more comprehensible by humans.

Traditionally, human experts or some semi-automatic mechanisms that incorporate human knowledge and computing techniques such as natural language processing provided these kinds of knowledge. For example, the MEDLINE corpus required considerable human effort to carry out categorization using a set of Medical Subject Headings (MeSH) categories (Mehnert, 1997). However, fully automatic generation of categories and their structure are difficult for two reasons. First, we need to select some important words as category terms (or *category themes*). We use these words to represent the themes of categories and to provide indexing information for the categorized documents. Generally, a category term contains only a single word or a phrase. The selection of the terms will affect the categorization result as well as the effectiveness of the categorization. A proper selection of a category term should be able to represent the general idea of the documents under the corresponding category. Such selections were always done by human linguistic experts because we need an insight of the underlying semantic structure of a language to make the selections. Unfortunately, such insight is hard to automate. Certain techniques such as word frequency counts may help, but it is the human experts who finally decide what terms are most discriminative and representative. Second, for the ease of human comprehension, the categories were always arranged in a tree-like hierarchical structure. This hierarchy reveals the relationships among categories. A category associated with higher-level nodes of the hierarchy represents a more general theme than those associated with lower level nodes. Also, a parent category in the hierarchy should represent the common theme of its child categories. The retrieval of documents of a particular interest can be effectively achieved through such hierarchy. Although the hierarchical structure is ideal for revealing the similarities among categories, the hierarchy must be constructed carefully such that irrelevant categories may not be the children of the same parent category. A thorough investigation of the semantic relations among category terms must be conducted to establish a well-organized hierarchy. This process is also hard to automate. Therefore, most of text categorization