Topic Maps Generation by Text Mining

Hsin-Chang Yang
Chang Jung University, Taiwan

Chung-Hong Lee
National Kaohsiung University of Applied Sciences, Taiwan

INTRODUCTION

Topic maps provide a general, powerful, and user-oriented way to navigate the information resources under consideration in any specific domain. A topic map provides a uniform framework that not only identifies important subjects from an entity of information resources and specifies the resources that are semantically related to a subject, but also explores the relations among these subjects. When a user needs to find some specific information on a pool of information resources, he or she only needs to examine the topic maps of this pool, select the topic that seems interesting, and the topic maps will display the information resources that are related to this topic, as well as its related topics. The user will also recognize the relationships among these topics and the roles they play in such relationships. With the help of the topic maps, you no longer have to browse through a set of hyperlinked documents and hope that you may eventually reach the information you need in a finite amount of time, while knowing nothing about where to start. You also don't have to gather some words and hope that they may perfectly symbolize the idea you're interested in, and be well-conceived by a search engine to obtain reasonable result. Topic maps provide a way to navigate and organize information, as well as create and maintain knowledge in an infoglut.

To construct a topic map for a set of information resources, human intervention is unavoidable at the present time. Human effort is needed in tasks such as selecting topics, identifying their occurrences, and revealing their associations. Such a need is acceptable only when the topic maps are used merely for navigation purposes and when the volume of the information resource is considerably small. However, a topic map should not only be a topic navigation map. The volume of the information resource under consideration is generally large enough to prevent the manual construction of topic maps. To expand the applicability of topic maps, some kind of automatic process should be involved during the construction of the maps. The degree of automation in such a construction process may vary for different users with different needs. One person may need only a friendly interface to automate the topic map authoring process, while another may try to automatically identify every component of a topic map for a set of information resources from the ground up. In this article, we recognize the importance of topic maps not only as a navigation tool but also as a desirable scheme for knowledge acquisition and representation. According to such recognition, we try to develop a scheme based on a proposed text-mining approach to automatically construct topic maps for a set of information resources. Our approach is the opposite of the navigation task performed by a topic map to obtain information. We extract knowledge from a corpus of documents to construct a topic map. Although currently the proposed approach cannot fully construct the topic maps automatically, our approach still seems promising in developing a fully automatic scheme for topic map construction.

BACKGROUND

Topic map standard (ISO, 2000) is an emerging standard, so few works are available about the subject. Most of the early works about topic maps focus on providing introductory materials (Ahmed, 2002; Pepper, 1999; Beird, 2000; Park & Hunting, 2002). Few of them are devoted to the automatic construction of topic maps. Two works that address this issue were reported in Rath (1999) and Moore (2000). Rath discussed a framework for automatic generation of topic maps according to a so-called topic map template and a set of generation rules. The structural information of topics is maintained in the template. To create the topic map, they used a generator to interpret the generation rules and

extract necessary information that fulfills the template. However, both the rules and the template are to be constructed explicitly and probably manually. Moore discussed topic map authoring and how software may support it. He argued that the automatic generation of topic maps is a useful first step in the construction of a production topic map. However, the real value of such a map comes through the involvement of people in the process. This argument is true if the knowledge that contained in the topic maps can only be obtained by human efforts. A fully automatic generation process is possible only when such knowledge may be discovered from the underlying set of information resources through an automated process, which is generally known as knowledge discovery from texts, or *text mining* (Hearst, 1999; Lee & Yang, 1999; Wang, 2003; Yang & Lee, 2000).

MAIN THRUST

We briefly describe the text-mining process and the generation process of topic maps in this section.

The Text-Mining Process

Before we can create topic maps, we first perform a text-mining process on the set of information resources to reveal the relationships among the information resources. Here, we only consider those information resources that can be represented in regular texts. Examples of such resources are Web pages, ordinary books, technical specifications, manuals, and so forth. The set of information resources is collectively known as *the corpus,* and individual resource is referred to as a *document* in the following text. To reveal the relationships between documents, the popular *self-organizing map (SOM)* algorithm (Kohonen, Kaski, Lagus, Salojärvi, Honkela, Paatero, & Saarela, 2000) is applied to the corpus to cluster documents. We adopt the vector space model (Baeza-Yates and Ribiero-Neto, 1999) to transform each document in the corpus into a binary vector. These document vectors are used as input to train the map. We then apply two kinds of labeling processes to the trained map and obtain two feature maps, namely the *document cluster map (DCM)* and the *word cluster map (WCM).* In the document cluster map, each neuron represents a document cluster that contains several similar documents with high word co-occurrence. In the word cluster map, each neuron represents a cluster of words revealing the general concept of the corresponding document cluster that is associated with the same neuron in the document cluster map.

The text-mining process described in the preceding paragraph provides a way for us to reveal the relationships between the topics of the documents. Here, we introduce a method to identify topics and the relationships between them. The method also arranges these topics in a hierarchical manner according to their relationships. As we mention earlier in this article, a neuron in the DCM represents a cluster of documents containing words that often co-occurred in these documents. Besides, documents that associate with neighboring neurons contain similar sets of words. Thus, we may construct a supercluster by combining neighboring neurons. To form a supercluster, we first define the distance between two clusters:

$$D(i,j) = H\left(\left\|\mathbf{G}_i - \mathbf{G}_j\right\|\right), \quad (1)$$

where i and j are the neuron indices of the two clusters, and $\mathbf{G}_i$ is the two-dimensional grid location of neuron i. $\|\mathbf{G}_i - \mathbf{G}_j\|$ measures the Euclidean distance between the two coordinates $\mathbf{G}_i$ and $\mathbf{G}_j$. $H(x)$ is a bell-shaped function that has a maximum value when x=0. We also define the dissimilarity between two clusters:

$$\delta(i,j) = \left\|\mathbf{w}_i - \mathbf{w}_j\right\|, \quad (2)$$

where $\mathbf{w}_i$ denotes the synaptic weight vector of neuron i. We may then compute the *supporting cluster similarity,* S_i, for a neuron i from its neighboring neurons by the equations

$$s(i,j) = \frac{\mathrm{doc}(i)\mathrm{doc}(j)}{F\left(D(i,j)\delta(i,j)\right)}$$
$$S_i = \sum_{j \in B_i} s(i,j) \quad (3)$$

where doc(i) is the number of documents associated with neuron i in the document cluster map, and B_i is the set of neuron indices in the neighborhood of neuron i. The function F: $\mathbf{R}^+ \rightarrow \mathbf{R}^+$ is a monotonically increasing function. A *dominating neuron* is the neuron that has locally maximal supporting cluster similarity.