Text Mining Methods for Hierarchical Document Indexing

Han-Joon Kim
The University of Seoul, Korea

INTRODUCTION

We have recently seen a tremendous growth in the volume of online text documents from networked resources such as the Internet, digital libraries, and company-wide intranets. One of the most common and successful methods of organizing such huge amounts of documents is to hierarchically categorize documents according to topic (Agrawal, Bayardo, & Srikant, 2000; Kim & Lee, 2003). The documents indexed according to a hierarchical structure (termed 'topic hierarchy' or 'taxonomy') are kept in internal categories as well as in leaf categories, in the sense that documents at a lower category have increasing specificity. Through the use of a topic hierarchy, users can quickly navigate to any portion of a document collection without being overwhelmed by a large document space. As is evident from the popularity of Web directories such as Yahoo (http://www.yahoo.com/) and Open Directory Project (http://dmoz.org/), topic hierarchies have increased in importance as a tool for organizing or browsing a large volume of electronic text documents.

Currently, the topic hierarchies maintained by most information systems are manually constructed and maintained by human editors. The topic hierarchy should be continuously subdivided to cope with the high rate of increase in the number of electronic documents. For example, the topic hierarchy of the Open Directory Project has now reached about 590,000 categories. However, manually maintaining the hierarchical structure incurs several problems. First, such a manual task is prohibitively costly as well as time-consuming. Until now, large search portals such as Yahoo have invested significant time and money into maintaining their taxonomy, but obviously they will not be able to keep up with the pace of growth and change in electronic documents through such manual activity. Moreover, for a dynamic networked resource (e.g., World Wide Web) that contains highly heterogeneous documents accompanied by frequent content changes, maintaining a 'good' hierarchy is fraught with difficulty, and oftentimes is beyond the human experts' capabilities. Lastly, since human editors' categorization decision is not only highly subjective but their subjectivity is also variable over time, it is difficult to maintain a reliable and consistent hierarchical structure. The above limitations require information systems that can provide intelligent organization capabilities with topic hierarchies. Related commercial systems include Northern Light Search Engine (http://www.northernlight.com/), Inktomi Directory Engine (http://www.inktomi.com/), and Semio Taxonomy (http://www.semio.com/), which enable a browsable Web directory to be automatically built. However, these systems did not address the (semi-)automatic evolving capabilities of organizational schemes and classification models at all. This is one of the reasons why the commercial taxonomy-based services do not tend to be as popular as their manually constructed counterparts, such as Yahoo.

BACKGROUND

In future systems, it will be necessary for users to be able to easily manipulate the hierarchical structure and the placement of documents within it (Aggarwal, Gates, & Yu, 1999; Agrawal, Bayardo, & Srikant, 2000). In this regard, this section presents three critical requirements for intelligent taxonomy construction, and taxonomy construction process using text-mining techniques.

Requirements for Intelligent Taxonomy Construction

(1) **Automated classification of text documents:** In order to organize a huge number of documents, it is essential to automatically assign incoming documents to an appropriate location on a predefined taxonomy. Recent approaches towards automated classification have used supervised machine-learning approaches to inductively build a classification model of pre-defined categories from a training set of labeled (pre-classified) data. Basically, such machine-learning based classification requires sufficiently large number of labeled training examples to build an accurate classification model. Assigning class labels to unlabeled documents should be performed by human labeler, and the task is a highly time-consuming and expensive. Furthermore, an online learning framework is nec-

Table 1. Procedure for hierarchically organizing text documents

Step 1. Initial construction of taxonomy
 i. Define an initial (seed) taxonomy
Step 2. Category (Re-) Learning
 i. Collect a set of the controlled training data fit for the defined (or refined) taxonomy
 ii. Generate (or Update) the current classification model so as to enable a classification task for newly generated categories
 iii. Periodically update the current classification model so as to constantly guarantee high degree of classification accuracy while refining the training data
Step 3. Automatic Classification
 i. Retrieve documents of interest from various sources
 ii. Assign each of the unknown documents into more than one categories with its maximal membership value according to the established model
Step 4. Evolution of taxonomy
 i. If concept drift or a change in the viewpoint occurs within a sub-taxonomy, reorganize the specified sub-taxonomy
 ii. If a new concept sprouts in the unclassified area, perform the cluster analysis for the data within the unclassified area into new categories
Step 5. Sub-taxonomy Construction and Integration
 i. Integrate the refined sub-taxonomy or new categories into the main taxonomy
Step 6. Go to Step 2

essary because it is impossible to distinguish training documents from unknown documents to be classified in the operational environment. In addition, classification models should be continuously updated so that their accuracy can be maintained at a high level. To resolve this problem, incremental learning methods are required, in which an established model can be updated incrementally without re-building it completely.

(2) **Semi-automatic management of evolving taxonomy:** The taxonomy initially constructed should change and adapt as its document collection continuously grows or users' needs change. When concept drift (which means that the general subject matter of information within a category may no longer suit the subject that best explained that information when it was originally created) happens in particular categories, or when the established criterion for classification alters with time as the content of the document collection changes, it should be possible for part of taxonomy to be reorganized; the system is expected to recommend users different feasible sub-taxonomies for that part.

(3) **Making use of domain (or human) knowledge in cluster analysis for topic discovery:** In order to refine the taxonomy, it is necessary to discover new topics (or categories) that can precisely describe the currently indexed document collection. In general, topic discovery is achieved by clustering techniques since clusters that are distinct groups of similar documents can be regarded as representing topically coherent topics in the collection. Clustering for topic discovery is a challenging problem with sufficient domain knowledge. This is because taxonomy should reflect the preferences of an individual user or specific requirements of an application. However, clustering is inherently an unsupervised learning process without depending on external knowledge. Therefore, a new type of supervised clustering is required that reflects external knowledge provided by users.

Taxonomy Construction Process using Text-Mining Techniques

Table 1 illustrates a procedure for hierarchically organizing text documents. The system begins with an initial topic hierarchy in which each document is assigned to its appropriate categories by automatic document classifiers. The topic hierarchy is then made to evolve so as to reflect the current contents and usage of indexed documents. The classification process repeats based on the more refined hierarchy.

In *Table 1*, steps 2 and 3 are related to machine-learning based text classification, step 4 semi-supervised clustering for topic discovery, and step 5 taxonomy building.

MAIN THRUST

This section discusses a series of text mining algorithms that can effectively support the taxonomy construction process. Recent text mining algorithms are prompted by machine learning paradigm; in particular, so are classification and clustering algorithms. Another important issue is about feature selection algorithms because textual data includes a huge number of features such as words or phrases. A feature selection module in the system extracts plain text from each of the retrieved documents and automatically determines only more significant features