Bridging Taxonomic Semantics to Accurate Hierarchical Classification

Lei Tang
Arizona State University, USA

Huan Liu
Arizona State University, USA

Jiangping Zhang
The MITRE Corporation, USA

INTRODUCTION

The unregulated and open nature of the Internet and the explosive growth of the Web create a pressing need to provide various services for content categorization. The hierarchical classification attempts to achieve both accurate classification and increased comprehensibility. It has also been shown in literature that hierarchical models outperform flat models in training efficiency, classification efficiency, and classification accuracy (Koller & Sahami, 1997; McCallum, Rosenfeld, Mitchell & Ng, 1998; Ruiz & Srinivasan ,1999; Dumais & Chen, 2000; Yang, Zhang & Kisiel, 2003; Cai & Hofmann, 2004; Liu, Yang, Wan, Zeng, Cheng & Ma, 2005). However, the quality of the taxonomy attracted little attention in past works. Actually, different taxonomies can result in differences in classification. So the quality of the taxonomy should be considered for real-world classifications. Even a semantically sound taxonomy does not necessarily lead to the intended classification performance (Tang, Zhang & Liu 2006). Therefore, it is desirable to construct or modify a hierarchy to better suit the hierarchical content classification task.

BACKGROUND

Hierarchical models rely on certain predefined content taxonomies. Content taxonomies are usually created for ease of content management or access, so semantically similar categories are grouped into a parent category. Usually, a subject expert or librarian is employed to organize the category labels into a hierarchy using some ontology information. However, such a taxonomy is often generated independent of data (e.g., documents). Hence, there may exist some inconsistency between the given taxonomy and data, leading to poor classification performance.

First, semantically similar categories may not be similar in lexical terms. Most content categorization algorithms are statistical algorithms based on the occurrences of lexical terms in content. Hence, a semantically sound hierarchy does not necessarily lead to the intended categorization result.

Second, even for the same set of categories, there could be different semantically sound taxonomies. Semantics does not guarantee a unique taxonomy. Different applications may need different category taxonomies. For example, sports teams may be grouped according to their locations such as Arizona, California, Oregon, etc and then the sports types such as football, basketball, etc.. Depending upon the application, they may also be grouped according to the sports types first and then locations. Both taxonomies are reasonable in terms of semantics. With a hierarchical classification model, however, the two taxonomies would likely result in different performances. Hence, we need to investigate the impact of different hierarchies (taxonomies) on classification.

In addition, semantics may change over time. For example, when the semantic taxonomy was first generated, people would not expect the category *Hurricane* related to *Politics*, and likely put it under *Geography*. However, after investigating the data recently collected, it is noticed that a good number of documents in category *Hurricane* are actually talking about the disasters Hurricane Katrina and Rita in the United States and the responsibility and the faults of FEMA during the crises. Based on the content, it is more reasonable to put

Hurricane under *Politics* for better classification. This example demonstrates the stagnant nature of taxonomy and the dynamic change of semantics reflected in data. It also motivates the data-driven adaptation of a given taxonomy in hierarchical classification.

MAIN FOCUS

In practice, semantics based taxonomies are always exploited for hierarchical classification. As the taxonomic semantics might not be compatible with specific data and applications and can be ambiguous in certain cases, the semantic taxonomy might lead hierarchical classifications astray. There are mainly two directions to obtain a taxonomy from which a good hierarchical model can be derived: *taxonomy generation via clustering* or *taxonomy adaptation via classification learning*.

Taxonomy Generation via Clustering

Some researchers propose to generate taxonomies from data for document management or classification. Note that the taxonomy generated here focus more on comprehensibility and accurate classification, rather than efficient storage and retrieval. Therefore, we omit the tree-type based index structures for high-dimensional data like R*-tree (Beckmann, Kriegel, Schneider & Seeger 1990), TV-tree (Lin, Jagadish & Faloutsos 1994), etc. Some researchers try to build a taxonomy with the aid of human experts (Zhang, Liu, Pan & Yang 2004, Gates, Teiken & Cheng 2005) whereas other works exploit some hierarchical clustering algorithms to automatically fulfill this task. Basically, there are two approaches for hierarchical clustering*: agglomerative* and *divisive*.

In Aggarwal, Gates & Yu (1999), Chuang & Chien (2004) and Li & Zhu (2005), all employ a hierarchical *agglomerative* clustering (HAC) approach. In Aggarwal, Gates & Yu (1999), the centroids of each class are used as the initial seeds and then projected clustering method is applied to build the hierarchy. During the process, a cluster with few documents is discarded. Thus, the taxonomy generated by this method may have different categories than predefined. The authors evaluated their generated taxonomies by some user study and found its perfomrance is comparable to the Yahoo directory. In Li & Zhu (2005), a linear discriminant projection is applied to the data first and then a hierarchical clustering method UPGMA (Jain & Dubes 1988) is exploited to generate a dendrogram which is a binary tree. For classification, the authors change the dendrogram to a two-level tree according to the cluster coherence, and hierarchical models yield classification improvement over flat models. But it is not sufficiently justified why a two-level tree should be adopted. Meanwhile, a similar approach, HAC+P was proposed by Chuang & Chien (2004). This approach adds one post-processing step to automatically change the binary tree obtained from HAC, to a wide tree with multiple children. However, in this process, some parameters have to be specified as the maximum depth of the tree, the minimum size of a cluster, and the cluster number preference at each level. These parameters make this approach rather ad hoc.

Comparatively, the work in Punera, Rajan & Ghosh (2005) falls into the category of *divisive* hierarchical clustering. The authors generate a taxonomy in which each node is associated with a list of categories. Each leaf node has only one category. This algorithm basically uses the centroids of the two most distant categories as the initial seeds and then applies Spherical K-Means (Dhillon, Mallela & Kumar, 2001) with k=2 to divide the cluster into 2 sub-clusters. Each category is assigned to one sub-cluster if majority of its documents belong to the sub-cluster (its ratio exceeds a predefined parameter). Otherwise, this category is associated to both sub-clusters. Another difference of this method from other HAC methods is that it generates a taxonomy with one category possibly occurring in multiple leaf nodes.

Taxonomy Adaptation via Classification Learning

Taxonomy clustering approach is appropriate if no taxonomy is provided at the initial stage. However, in reality, a human-provided semantic taxonomy is almost always available. Rather than "start from scratch", Tang, Zhang & Liu (2006) proposes to adapt the predefined taxonomy according the classification result on the data.

Three elementary hierarchy adjusting operations are defined:

- **Promote:** Roll up one node to upper level;
- **Demote:** Push down one node to its sibling;