Hierarchical Text Classification

Massand Sagar Sunil (11414)
Y. Kushal (11825)

February 28, 2014

1 Introduction

Hierarchies are important for the classification of documents on the Internet . With the exponential increase in the number of documents available online, there is a need for automated classification of new documents to the categories in a hierarchy. As more documents become available, more categories are also added to the hierarchy, so novel machine learning algorithms are required to solve the classification problem.

2 Summary of previous work

There are several methods available for such problems in the literature. Most of these approaches fall into one of the following categories -

Flat Classification Methods -
This basically ignores the hierarchy of the categories. These include the k-NN method, naive bayes technique etc. These work well for smaller number of categories but perform poorly for large number of categories as they fail to take into account the hierarchical structure of the categories.

Top-Down Classification Methods -

These methods use a set of local classifiers to distinguish between categories at each level of the hierarchy tree. Example - a hierarchical SVM(one for each internal category).

The disadvantage of these methods is that if there a misclassification at a parent category, the document is mis-classified in the subsequent levels as well.

Big-Bang Classification Methods -

This is also known as the global classifier approach. In this approach, a global classification model is used in contrast to the local classifiers in the top-down approach. The advantage of this method is that dependencies between each of the classes can be taken into account easily.

This does not work well when the categories are large in number.

Deep Classification Methods -

This method assumes that the subset of categories relevant to a document is sufficiently small. Firstly, a set of similar categories is obtained(searched for) from the large set of categories. In some sense, we are pruning the hierarchy tree for relevant categories.

Then, using the candidate categories we use one of the previous methods to classify/label the document.

3 Datasets

We use a dataset created from Wikipedia for a challenge based on a similar problem on Kaggle. The dataset is multi-class, multi-label and hierarchical. We plan to use a small subset of the categories in the dataset.

4 References

- Kaggle Challenge - `http://www.kaggle.com/c/lshtc/`
- *A survey of hierarchical classification across different application domains*, authored by Carlos N. Silla Jr., Alex A. Freitas
- `http://research.microsoft.com/enus/events/internetservices2008/guirong_xue.pdf`