Distributed Categorizer for Large Category Systems

Eric GAUSSIER a and Francois PACULL b,1

a University Joseph Fourier, Grenoble, France
b Xerox Research Centre Europe, Meylan, France

Abstract. We describe here a categorization system that allows to manage an arbitrary number of categories. It is based on a cascade of categorizers arranged as a tree. Categorizers are deployed on a set of machines to parallelize the processing.

Keywords. Categorizer, Large Scale System.

Introduction

The memory and time complexity is a crucial problem for categorization algorithms. Indeed, the complexity is somehow linked to the number of considered classes multiplied by the size of the vocabulary used during the training phase. Managing a large category system with a single standalone categorizer is most often impossible since these two parameters increase at the same time as the category system grows.

The solution proposed here allows to manage an arbitrary number of categories. It is based on a cascade of categorizers arranged as a tree. Each node of the tree is a categorizer the role of which is either to find the next categorizer(s) (for the intermediary categorizers) or the actual categories (for the final categorizers). From a practical point of view, the intermediary categorizers act as coarse-grained switches while the final categorizers play their classical roles and return the possible matching categories with the corresponding probabilities. These probabilities are then aggregated and unified to define the final result.

1. Definition of the category system

In figure 1-a, we considered a system based on a significant part of the Dmoz Open Directory Project (http://www.dmoz.org): namely the categories corresponding to the topics Arts, Business, Home, Reference, Science, Computers, News, Regional, Society, Health, Recreation and Sports. For the decomposition into intermediary categorizers, we first tried the most intuitive that consists of considering the set of immediate sub-categories as shown in the figure 1-b. However, when categorizers of a level are not homogeneous in term of sub-categories and vocabulary, including further sub-categories has shown to improve the results.

1 Corresponding Author: Francois.Pacull@xtce.xerox.com
2. Cascading processing and aggregation of the results

The cascade approach allows several algorithms and strategies to define at each intermediary level which paths have to be followed. For instance, we can consider to follow several paths in parallel if we want to process documents containing different aspects (e.g. Computer and Business or Sciences and Health). Alternatively, we can decide to really focus on the main category and then to follow a direct path to a single final categorizer.

Since the results of the categorization returned by final categorizers are independent we have to aggregate them. To do so, we have used calibration [1] to re-rank the different results in order to provide an homogeneous result across the whole category system.

Conclusion

The main advantages of our architecture are: (i) it uses a classical categorizer (Probabilistic Latent categorizer [2]) as a black box for the nodes of the systems: it is not required to modify the code of the categorizer; (ii) it allows to consider a very large number of categories: we have successfully experimented systems based on the United States Patent and Trademark (USPTO http://www.uspto.gov) category system (15 000 categories) and a larger one (100 000 categories) based on Dmoz. (iii) the cascading approach is flexible enough to really cope with the desired results and the specificity of the category system. (iv) it is possible to put in place a very efficient and low cost system: because the different categorizers do not require a lot of resources and are independent, the whole system can be distributed on a set of basic desktops. Then processing several times the same document (up to 5 in practice) is highly compensated by the pipelining and the parallelisation of the processing. We reached almost six hundred documents per minute with the system of 100 000 categories deployed on just 5 machines.

References