Chapter XII
A Hierarchical Online Classifier for Patent Categorization

Domonkos Tikk
Budapest University of Technology and Economics, Hungary

György Biró
TextMiner Ltd., Hungary

Attila Törcsvári
Arcanum Development Ltd., Hungary

ABSTRACT

Patent categorization (PC) is a typical application area of text categorization (TC). TC can be applied in different scenarios at the work of patent offices depending on at what stage the categorization is needed. This is a challenging field for TC algorithms, since the applications have to deal simultaneously with a large number of categories (in the magnitude of 1,000–10,000) organized in hierarchy, large number of long documents with huge vocabularies at training, and they are required to work fast and accurate at on-the-fly categorization. In this chapter we present a hierarchical online classifier, called HITEC, which meets the above requirements. The novelty of the method lies in the taxonomy dependent architecture of the classifier, the applied weight updating scheme, and in the relaxed category selection method. We evaluate the method on two large English patent application databases, the WIPO-alpha and the Espace A/B corpora. We also compare the presented method to other TC algorithms on these collections and show that it outperforms them significantly.
INTRODUCTION

The immense growth in the number of electronic documents necessitates powerful algorithms and tools that are able to manage data of such quantity. An obvious way to handle the vast number of documents is organizing them according to their topic into hierarchical category systems (taxonomies), which eases the search, retrieval, insertion, and storage. In general, the set-up of such document-management systems from scratch requires considerable time and cost, both at the time when the taxonomy is created and when it is filled up. In this chapter, we concentrate on the field of patent categorization where these initial steps can be skipped, because various taxonomies and patent databases are already available.

The rapid growth of data can also be observed in this field: the number of patent applications is increasing every year, demanding higher and higher level of computer support for categorization-related tasks. Let us now briefly mention some properties of patent categorization (PC) problems that are challenging for text categorization (TC) algorithms.

Patent offices organize patent applications into very large topic taxonomies. The most important among them is the international patent classification (IPC), which is a standard taxonomy developed and maintained by the World Intellectual Property Organization (WIPO), but other national taxonomies are also in use, such as the US patent Classification (UPC), the Canadian Patent Classification (CPC), and others. The IPC consists of about 80,000 categories that cover the whole range of industrial technologies. There are 8 sections at the highest level of the hierarchy, then 128 classes, 648 subclasses, about 7,200 main groups, and about 72,000 subgroups at lower levels. From the total 80,000 categories, the top four levels are mostly used in automated patent categorization systems.

Documents are also quite large compared to usual benchmark domains of TC such as, for example, business news (Reuters-21578 and its variants, Reuters Corpus Volume 1), and short communications (20 Newsgroups data). An average document contains more than 3,000 words. A patent application document consists of several fields, annotated by metadata, a feature which may be partly or fully taken into account at categorization.

The vocabulary of patent applications is quite diverse, which leads to an extremely large dictionary. Many vague or general terms are often used in order to avoid narrowing the scope of the invention (Fall, Torcsvari, Benzineb & Karetka, 2003). Combinations of general terms often have a special meaning that also has to be captured. Patent documents also include acronyms and much new terminology (Kando, 2000). Moreover, unlike news stories, patents are necessarily all different at a semantic level, as each must describe a new invention. This may complicate the training of a classifier. Furthermore, IPC categories often cover large and disparate areas, resulting in large vocabularies in even in single categories.

The patent applications are usually assigned to one primary category and several secondary categories. At the evaluation of TC approaches, this characteristic has to be also taken into account. For this task, particular evaluation functions were proposed (Fall, Torcsvari, Benzineb et al., 2003; see also the experimental section).

Therefore TC approaches for PC problems have to manage simultaneously:

- A very large size of taxonomy,
- Large documents,
- A huge feature set,
- Multilabeled documents.

Finally, another often occurring phenomenon of patent classification is that capturing the semantics of a document statistically is not sufficient for proper categorization. This can be explained by the nature of how the primary categories are assigned to documents: the primary category is