Chapter 13

EXPERIMENT WITH A HIERARCHICAL TEXT CATEGORIZATION METHOD ON WIPO PATENT COLLECTIONS

Domonkos Tikk, György Biró, and Jae Dong Yang

1. INTRODUCTION

The immense and exponentially growth in the number of electronic documents stored on the internet, corporate intranets and data warehouses necessitates powerful algorithms and tools that are able to deal with data of such quantity. An obvious way to handle the vast number of documents is organizing them into category systems. Category systems are usually hierarchic (called taxonomy) because that offers straightforward way to find and browse data at arbitrary refinement. E.g. documents on large internet directories, such as Yahoo! and Google, are categorized into taxonomy. This storage technique requires efficient automatic categorization methods as manual text categorization is no longer amenable in that size, requiring a vast amount of time and cost.

The purpose in automatic text categorization (TC) is to assign a document to appropriate category/ies (or topic) being selected from a predefined set of categories. Originally, research in TC addressed the binary problem, where a document is either relevant or not w.r.t. a given category. In real-world situation, however, the great variety of different sources and hence categories usually poses multi-class classification problem, where a document belongs to exactly one category selected from a predefined set Baker and McCallum, 1998; Weiss et al., 1999; Wiener et al., 1993; Yang, 1999. Even more general is the case of multi-label problem, where a document can be classified into more than one category. While binary and multi-class problems were investigated extensively Sebastiani, 2002, multi-label problems have received much less attention Aas and Eikvil, 1999.

As the number of topics becomes larger, multi-class categorizers face the problem of complexity that may incur rapid increase of time and storage, and compromise the perspicuity of categorized subject domain. A common way to manage complexity is using a hierarchy (in this paper we restrict our investigation to tree structured hierarchies), and text is no exception Chakrabarti et al., 1998. Internet directories (see e.g. Yahoo; http://www.yahoo.com) and large on-line databases are often organized in hierarchy.

Patent databases are typically such where the use of a hierarchical category system is a necessity. Patents cover a very wide area of topics, and each field can be further divided into subtopics, until a reasonable level of specialization is reached. The International Patent Classification (IPC) is a standard taxonomy developed and administered by WIPO (World Intellectual Property Organization) for classifying patents and patent applications. The use of patent documents and IPC for research into automated categorization is interesting for the following reasons Fall et al., 2002:

1 IPC covers a huge range of topics and uses a diverse technical and scientific vocabulary.

2 IPC is a complex, hierarchical taxonomy, where over 40 million documents have been classified worldwide. The number of documents classified each year is rising fast.

3 Domain experts in national patent offices currently classify patent documents fully manually. These experts have an intimate knowledge of the IPC system.

4 Patent documents are often available in several languages. Professional translators have already performed large numbers of translations manually.

As a courtesy of WIPO, we could experiment with the WIPO-alpha English and WIPO-de German patent databases issued in late 2002 and early 2003, respectively. (Collections are available after registration at http://www.wipo.int/ibis/datasets/index.html.) WIPO-alpha is a large collection (3 GB) of about 75000 XML documents distributed over 5000 categories in four levels (the top four levels of IPC); WIPO-de is an even larger collection of about 110000 XML documents defined on the same taxonomy (IPC). At WIPO, they experimented with several text categorization technique on the WIPO-alpha collection, see Fall et al., 2003a.