Information Processing and Management 48 (2012) 1017–1030

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A three-phase method for patent classification

Yen-Liang Chen *, Yuan-Che Chang

Department of Information Management, National Central University, Chung-Li 320, Taiwan, ROC

ARTICLE INFO

Article history:
Received 13 January 2011
Received in revised form 21 November 2011
Accepted 29 November 2011
Available online 9 January 2012

Keywords:
Patent classification
Vector space model (VSM)
IPC taxonomy
Support vector machines (SVM)
K-means
K nearest neighbors (KNN)

ABSTRACT

An automatic patent categorization system would be invaluable to individual inventors and patent attorneys, saving them time and effort by quickly identifying conflicts with existing patents. In recent years, it has become more and more common to classify all patent documents using the International Patent Classification (IPC), a complex hierarchical classification system comprised of eight sections, 128 classes, 648 subclasses, about 7200 main groups, and approximately 72,000 subgroups. So far, however, no patent categorization method has been developed that can classify patents down to the subgroup level (the bottom level of the IPC). Therefore, this paper presents a novel categorization method, the three phase categorization (TPC) algorithm, which classifies patents down to the subgroup level with reasonable accuracy. The experimental results for the TPC algorithm, using the WIPO-alpha collection, indicate that our classification method can achieve 36.07% accuracy at the subgroup level. This is approximately a 25,764-fold improvement over a random guess.

1. Introduction

In recent years, the number of patent documents has increased rapidly, thereby increasing the demand for powerful algorithms and systems for automatic categorization tasks (Kim & Choi, 2007). Patents are structured to include the claims, purpose, effects, embodiments, etc. of the invention. To improve patent management, search, and retrieval, patent offices all around the world must assign classification codes to each patent application so that patents with similar characteristics can be placed in the same subdirectory. Therefore, it is not surprising that different classification hierarchies have been proposed by different patent offices worldwide. Usually, each national patent office uses its own hierarchy, such as the United States Patent Classification (USPC) in the United States, the European Classification (ECLA) for the European Union, the File Index (FI) for Japan, and so on.

Among these classification systems, the most important is the IPC, a complex hierarchical classification system comprised of sections, classes, subclasses, and groups, which is a standard taxonomy developed and maintained by the World Intellectual Property Organization (WIPO) for classifying patents and patent applications. The IPC covers all areas of technology and is currently used by industrial property offices in more than 90 countries (Fall, Törcsvári, Fievét, & Karetka, 2004). It includes some 80,000 categories that cover the whole range of industrial technologies. Eight sections (A through H) are at the highest level of the hierarchy, followed by 128 classes, 648 subclasses, about 7200 main groups, and approximately 72,000 subgroups at the bottom level (Tikk, Biró, & Törcsvári, 2007).

Research regarding automated categorization with patent documents and the IPC is interesting (WIPO (2 May 2011)), especially since it has become more common worldwide to classify all patent documents using the IPC structure and code. As a result, many researchers have begun developing automatic methods or systems to help users assign patent the IPC codes. These methods can be roughly classified into two major approaches: single level methods and hierarchical methods.

* Corresponding author.
E-mail address: ylchen@mgt.ncu.edu.tw (Y.-L. Chen).

0306-4573/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.11.001

Table 1
Wimbledon Championships for men's singles (2001–2010). *Source*: http://en.wikipedia.org.

Year	Champion	Previous year end ranking
2001	Goran Ivanisevic	129
2002	Lleyton Hewitt	1
2003	Roger Federer	6
2004	Roger Federer	2
2005	Roger Federer	1
2006	Roger Federer	1
2007	Roger Federer	1
2008	Rafael Nadal	2
2009	Roger Federer	2
2010	Rafael Nadal	2

Single level methods regard patents as texts and employ traditional classification methods to assign the IPC codes. Methods that have been used include Naive Bayes (NB), KNN, SVM algorithms, and a variant of Winnow (Fall et al., 2004). A common drawback, however, of these flat text classifiers is that since they treat patents as texts without using the patents' hierarchical structural properties to further enhance classification accuracy, the classification results do not appear to be so satisfactory. According to test results with the WIPO-alpha collection by Fall et al. (2004), NB and SVM performed the best (55%) at the class level, while SVM outperformed other methods (41%) at the subclass level.

Intuitively, these figures do not look so good. Since there are 128 classes and 648 subclasses in the IPC hierarchy, however, they are in actuality much better than those obtained by random guessing, which are 1/128 and 1/648 respectively. In other words, although the classification results are not completely satisfactory, it may be difficult to improve their classification accuracy. This also explains why no single level classifier has yet been used to classify codes below the subclass level.

Because of the single level methods' weaknesses, another approach, which focuses on classifying code using the patents' hierarchical properties, is utilized. To the best of our knowledge, the best work in this area is by Tikk et al. (2007), who proposed an automatic patent categorization method, called HITEC, which classifies the code hierarchically, level by level, according to the IPC hierarchical structure. The results obtained with HITEC using data from the WIPO-alpha corpus show that it achieved 53.25% accuracy at the subclass level, which is 12% more accurate than the single level classifiers, and it achieved an accuracy of 36.89% at the main group level.

Interestingly, the work of Tikk et al. (2007) only classified code at the main group level. No accuracy results were reported at the subgroup level. In the following, we discuss possible difficulties that may occur if we employ the hierarchical approach to classifying code at the bottom level, which contains about 72,000 subgroups.

Since hierarchical methods classify code level by level, from the top to down, the classification accuracy declines continuously as more levels are traced. Suppose that the classification accuracy of each level is about 85%. Then, since we have five levels in the hierarchy, the bottom level's expected accuracy would be 44%. Unfortunately, in reality it may be even lower because the lower level usually has a much larger number of branches. Additionally, the texts in different branches at lower levels are more difficult to distinguish than those in different branches at higher levels. These difficulties explain why no hierarchical classifier has been used to classify code below the main group level.

Since neither the single level methods nor the hierarchical methods can classify code at the bottom level, the question is if there are any more patent properties that can be used to improve accuracy so that IPC code can be classified to the bottom level. This has motivated research into what patent properties can be used to further improve classification accuracy and how to use them. In this study, our goal is to investigate one possible answer to this question. Below, we discuss two important observations that make this possible.

- Observation 1:
  (1) The first observation is that the bottom level of the IPC code contains numerous categories (72,000 subgroups), so it is very difficult to determine the target patent's category using either single level methods or hierarchical methods. The probability of making a correct guess, however, increases sharply if the requirements are relaxed a little. For example, the list of Wimbledon Champions for men's singles from 2001 to 2010 is shown in Table 1. If we guess the Wimbledon Champion every year using the top athlete from the previous year-end ATP (Association of Tennis Professionals) ranking, then the average prediction accuracy is 40%. The prediction accuracy can be greatly improved, however, if we predict which ten people may win the Wimbledon Championships. By predicting the winner every year as one of the top 10 people in the previous year-end ranking, the average accuracy increases to 90%.
  (2) Based on this idea, we can first use classifiers to obtain the top k_1 subclasses at the subclass level, and then obtain the top k_2 subgroups from the subgroups in these k_1 subclasses. Assume that $k_1 = 10$, $k_2 = 20$ and the classification accuracy at every level, according to our experiment results, is about 87%. Then, we can reduce the number of possible subgroups from 72,000 to 20 and keep the accuracy at about 75%.