Applied Soft Computing 10 (2010) 1164–1177

Contents lists available at ScienceDirect

Applied Soft Computing

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

Patent classification system using a new hybrid genetic algorithm support vector machine

Chih-Hung Wu[a,*], Yun Ken[b], Tao Huang[b]

[a] Department of Digital Content and Technology, National Taichung University, 140, Min-Shen Road, Taichung 40306, Taiwan, ROC
[b] Department of Business Administration, National Yunlin University of Science and Technology, Taiwan, ROC

ARTICLE INFO

Article history:
Received 3 May 2008
Received in revised form 3 November 2009
Accepted 27 November 2009
Available online 24 December 2009

Keywords:
Patent classification
Hybrid genetic-based support vector machine (HGA-SVM)
Parameter optimization

ABSTRACT

Nowadays, decision-making activities of knowledge-intensive enterprises depend heavily on the successful classification of patents. A considerable amount of time is required to achieve successful classification because of the complexity associated with patent information and of the large number of potential patents. Several different patent classification approaches have been developed in the past, but most of these studies focus on using computational models for the International Patent Classification (IPC) system rather than using these models in real-world cases of patent classification. In contrast to previous studies that combined algorithms and the IPC system directly without using expert screening, this study proposes a novel artificial intelligence (AI)-aided patent decision-making process. In this process, an expert screening approach is integrated with a hybrid genetic-based support vector machine (HGA-SVM) model for developing a patent classification system with the high classification accuracy and generalization ability for real-world patent searching cases. The proposed approach is tested on a real-world case—an expert's patent document searching history that contains 234 patent documents of semiconductor equipment components. The research results demonstrate that our proposed hybrid genetic algorithm approach can optimize all the parameters of the SVM for developing a patent classification system with a high accuracy. The proposed HGA-SVM model is able to dynamically and automatically classify patent documents by recording and learning the experts' knowledge and logic. Finally, we propose a new decision-making process for improving the development of the SVM patent classification and searching system.

1. Introduction

Obtaining high-quality patent information is one of the critical factors to ensure success in all the steps of the business decision-making process, from creating an idea to launching a successful product on the market. Everyone along the chain of innovation, from research scientists to marketing managers, is in some way affected by existing or potentially granted patents [1]. Consequently, it is difficult for patent information specialists, patent attorneys, database producers, and patent offices to ensure the high-quality patent information because of the requirements of thoroughness, timeliness, and accuracy.

Unfortunately, the increasing volume of information makes it is almost impossible to efficiently determine critical patent information in the business decision-making process. Approximately one million patent applications have been published worldwide each year. Many patent documents describe a small improvement upon the state-of-the-art. Many "doubtful" applications are being filed and granted, thereby increasing the workload and reducing efficiency [1]. The enormous increase in the number of patent applications is creating significant challenges for the entire patent system and all of the patent information users. In this situation, the immediate problem is to find a method to cope with the enormous amount of patent information being generated. Prior to the advent of the Internet, professional users assumed that information retrieved from electronic storage was a faithful reflection of their search strategy. The implementation of paid-for links on Internet search engines has demolished that assumption. We can no longer assume that the results presented to us as "the most relevant" are the best fit to our strategy. Therefore, an efficient patent classification system that can help professional users in patent searching and provide highly accurate results plays a critical role in solving the abovementioned problem.

Text categorization techniques can classify texts (original language) into various categories defined in advance by the researcher [2]. The increasingly rapid changes in information technology and the considerable increase in Internet usage have turned all the available information into a huge mass of jumbled data. Consequently, sorting through large amounts of data has become a very

* Corresponding author. Tel.: +886 422183024; fax: +886 422183270.
E-mail addresses: chwu@ntcu.edu.tw (C.-H. Wu), yunken@yuntech.edu.tw (Y. Ken), g9522808@yuntech.edu.tw (T. Huang).

1568-4946/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2009.11.033

important task. Because of this, many automatic algorithms have been developed in recent years, such as the support vector machine (SVM) [3–7], Naïve Bayes [3–6,8], artificial neural network [9–12], K-nearest neighbor [6,13,14], and scene labeling using a relaxation algorithm [13]. Among them, the SVM is one of the more recent ones in calculus technology. Further, the research of Joachims [4] indicated that it has the advantages of "a high dimensional input space," "less unimportant characteristic values," "sparse vectors among marking files," "the fact that most of the text categorization problems are linear and divisible," etc. Many researches have proved that this method does have outstanding classification validity [3,4,15,16].

The most widely used patent classification system is the International Patent Classification (IPC) system established by the Strasbourg Agreement in 1971. It is used to arrange, retrieve, and manage patent data. This method has been used in more than 50 countries. In recent years, people have begun to pay more attention to the concept of patents. This move is also reflected in the changes in the number of patent applications from different countries. The number of patent applications received by the United States Patent and Trademark Office (USPTO) in 1997 amounted to 240,000, reaching approximately 380,000 in 2005 (http://www.USPTO.gov). The European Patent Office (EPO) received approximately 100,000 patent applications in 1997; this number increased to 180,000 in 2004 (http://www.European-patent-office.org). With this rapid increase in the number of patent applications, the workload of the patent sorters also increased. If computer technology could be used to deal automatically with the technical aspects of the patent classification process, the efficiency and accuracy of patent classification would be greatly improved. To provide an example, to ensure high-quality patent data, the EPO provides users with online patent data from more than 80 different countries and monitors the three main patent data streams (bibliographic, facsimile images, and full text) at key milestones throughout the data life cycle in order to ensure that the data are complete, consistent, accurate, and up to date.

In order to automatically classify patents, the patent divisions of many countries began related work, for example, USPTO [17,18], EPO [19], Japanese Patent Office [20,21], Bureau of Patent in France [22], and Xerox [23]. However, the automatic patent classification system was only used in the patent divisions of major countries, mainly as a research project; it was seldom used in practice [24].

A large majority of previously published studies [24–26] on this subject are based on the IPC system. However, the problem is that when a patent application is considered or submitted, the search for previous inventions in the field is facilitated by a prior accurate patent classification [24]. Therefore, in contrast to previous studies that combined algorithms and the IPC system directly without using expert screening, this study proposes a novel artificial intelligence (AI)-aided patent decision-making process by integrating an expert screening approach and a hybrid genetic-based SVM (HGA-SVM) model for developing a patent classification system with a high classification accuracy and generalization ability on real-world patent searching cases.

2. Review of patent classification and methodologies

This section reviews patent classification and methodologies. In the first section, we address the benefit of having a patent classification system in the business decision-making process. In the next section, we introduce the concepts of patent classification and patent searching. In the third section, we discuss the basic idea, theory, and concepts of the SVM for developing the patent classification system. Finally, we introduce the indexes that are used in this study for evaluating the accuracy of the proposed classification system.

2.1. Benefit of patent classification system for decision-making process

Using a patent classification system for patent analysis can improve the quality of patent information and facilitate the decision-making process while lowering the possibility of patent infringement. A good patent classification system can be used for improving the quality of patent information by identifying and classifying critical patent documents. High-quality patent information can be used for assisting the decision-making process in many areas—competitor monitoring, technology assessment, R&D portfolio management, human resource management, and the identification and assessment of potential sources for the external generation of technological knowledge, especially by means of mergers and acquisitions [27].

Strategic patent information is beneficial for two important recipients: (1) senior managers, who use this information for assisting the decision-making process in important areas such as technology management, and (2) external stakeholders of the firm, such as shareholders and analysts, who have an increasing interest in technological competence assessment because of its strong effect on the firm's future competitiveness [27].

Patents contain rich strategic technology information. Intelligent exploitation of patent information, a unique source of technical, business, and legal information, will contribute to the success of large or small enterprises. Many areas of business can benefit from patent information analysis. Some of the practical applications include: providing input to licensing strategy, cross-licensing, supporting mergers and acquisitions, guiding R&D, providing good indicators of the R&D output, facilitating human resource allocation (brain maps can identify star inventors within a company and in other companies, providing a valuable tool for retaining or head-hunting talented individuals), and serving as a tool for creative thinking [28].

Making a right decision related to licensing strategy requires patent information analysis to provide firms with reliable information, data on the availability of competing technologies in the same field, and a good understanding of the value of the target technology. While preparing to license-in technology, a company should analyze patent information via a patent searching or classification system in order to determine whether the technology in question is protected or is already in the public domain in the target market because of non-protection, expiration, non-payment of maintenance fees, or invalidation of the patent in a court proceeding; otherwise, the owner of an existing or competing patent could initiate an action for infringement. Further, such a system could help in determining whether the technology is overvalued or undervalued in comparison with other related or alternative technologies [28].

Cross-licensing is an agreement between two companies to license one or more patents to each other; it may involve payments if one of the parties is perceived to have the patent portfolio of a lesser value than the other. In order to making a successful cross-licensing decision, the firms usually adopt patent analysis for determining the optimal patent portfolios of the two companies.

The firms should first identify all the companies with relevant patents while acquiring a specific technology along with other complementary assets through a merger or acquisition. Further, a patent classification/search system will help to narrow the range of choices and decide which company is the best target.

In order to guide the R&D output, a company should use a patent classification/search system and patent information analysis to study the overview of the relevant technological field in order to forecast market needs. By using a patent search and classification