Automatic Patent Classification by a Three-Phase Model with Document Frequency Matrix and Boosted Tree

Fatima Al Shamsi
Department of Electrical Engineering and Computer Science
Masdar Institute of Science and Technology
Abu Dhabi, UAE
e-mail: fsalshamsi@masdar.ac.ae

Zeyar Aung
Department of Electrical Engineering and Computer Science
Masdar Institute of Science and Technology
Abu Dhabi, UAE
e-mail: zaung@masdar.ac.ae

***Abstract*—With the increased volume of patent databases during the past years, it becomes necessary for companies to correctly classify and identify innovative patents in a timely manner though the use of automation. Although many patent classification methods have been proposed, the accuracy remains the most challenging factor for the success of a classification model. This paper presents an empirical study for automatic patent classification systems through the application of a three-phase model. Patent query, text processing, and the classification phases are applied, and a document frequency matrix and boosted tree (BT) classifier are used to classify patents into two classes. Model validation, accuracy and performance are calculated to determine the effectiveness of the proposed model.**

***Index Terms*—Patent analysis, patent automation, machine learning, classification**

I. Introduction

Companies nowadays are motivated to identify patents that are worth investing in to achieve maximum profit gain or for guidance when developing new systems. Nevertheless, the constant increase in patents makes it critical to speed up the classification process when searching for new innovative systems in patent databases. The research on automatic patent classification and filtering during the recent years is because of the need for a cost- and time-effective technique to support system development and innovation processes.

Technology-specific patents are often purchased by companies to improve their own technologies and/or create new products. Another recent scenario is whereby manufacturers are charged of violating intellectual property by their competitors which prevents new products from entering the global market [1]. Alternatively, companies often require patent classification and analysis to inspire ideas when developing new systems. Conducting patent analysis using traditional techniques is very inefficient in terms of time, cost and manpower [2].

Boosted tree (BT) classifier is an affective learning technique that has been applied to numerous low-dimensional applications. The main functionality is to achieve a maximum correlation of new learners with the negative gradient of the loss function based on earlier iterations of the learning scheme [3]. BT trains many weak classifiers from the input data and then combine all the resulting classifiers into a single tree [4]. Any classifier that performs better than random guess can be used a weak learner [5].

In this study, we propose an automatic patent classification system using a three-phase model. The first phase is based on patent query, whereas the second phase focuses on text processing and producing the document frequency matrix with the results. The final phase is the classification model, where the boosted tree (BT) classifier, automatic classification, model validation and performance measurement are applied.

This paper is structured as follows. Section II presents a review of the literature that includes background information on patent classification and some prior related works. Section III discusses the methodology of our proposed scheme. Section IV presents experimental results and evaluation. Section V presents concludes the paper and mentions future work.

II. Literature Review

A. Patent classification

The most common method for patent classification is using the International Patent Classification (IPC) system which is managed by the World Intellectual Property Organization (WIPO). The patent classification systems is designed by leading countries with huge patent databases, including USA (USPTO), Europe (EPO), and Japan (JPO) [6]. The levels of classification represent an index containing classes, subclasses, groups, and subgroups [7].

An alternative method is the Association Rule-based Text Classification (ARTC). Associative text categorization defines strong rules with an association with class labels of which then benefits from generated rules to build classifiers for new objects. All testing documents are scanned for the search of association rules of which are generated from the training documents and afterward given a rank which is generally equivalent to the total weight of the rules found in the new record. Then, each document is allocated to a class if the ranking score is higher than the required threshold [8].

978-1-5090-5306-3/16/$31.00 ©2016 IEEE

Single-level methods view patents as plain text and performs basic classification to specify the International Patent Classification (IPC) codes. Methods such as Naive Bayes (NB), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) algorithms are widely used. Nevertheless, a common disadvantage of applying this type of basic text classifiers is that they generally view patents as text without using the patents' hierarchical structural properties to enhance classification accuracy [9].

B. Related works

Chen and Chang [10] presented a three-phase method for patent classification by dividing patents into subgroups and thus classifying the bottom-level of the International Patent Classification (IPC) hierarchy. They first start by merging flat text classifiers at two different hierarchy level with a clustering technique, and then apply TF-ICF to select discriminative terms within categories. Moreover, they apply the K-means algorithm to cluster all patents in the same subgroup, and then use the KNN algorithm with cosine similarity measures to determine the final IPC subgroup category.

Wu et al. [2] proposed a patent quality analysis and classification system using self-organizing maps with support vector machine. They clustered patents into different quality groups using self-organizing maps. Kernel principal component analysis was conducted to improve classification performance, and the support vector machine was used to build a strong classification model. They conduced an experiment to classify patent quality of thin film solar cells in solar power industry.

Eito-Brun [11] performed a case study on automatic patent classification techniques. They conducted citation analysis on leading industrial organizations in innovation to identify the transfer of technical knowledge between the organizations participating in the innovation activities. Analysis of productivity summarizing the most productive organization in terms of the total number of patent citations and the impact analysis by determining the most influencing organization were carried out by comparing the impact on subsequent researches.

Wu, Ken, and Huang [6] emphasized on using a new hybrid genetic algorithm support vector machine (SVM) approach. Their research resulted in a system that automatically identifies critical keywords obtained from several sections of patent documents for correctly classifying patents. They combined an expert screening approach and the SVM algorithm for developing the patent classification system.

III. Proposed Model

The proposed model follows the three phases as described below:

A. Phase 1: Patent query

- **Step 1: Collect patent data manually**
  Patents are collected manually from patent databases such as Google Patent Search and FreePatentsOnline. It is a basic search depending on keywords of interest.
- **Step 2: Calculate keyword frequency and query patents using patent keyword dictionary**
  After collecting the patents manually, keyword frequencies are calculated to generate a keyword dictionary which is then used to query new related patents which allows us to enrich the patent database.

B. Phase 2: Text processing

- **Steps 1 and 2: Choose abstract and remove stop words**
  An abstract is chosen as the variable of interest. Stop words are removed from the variable of interest to ensure the accuracy of the following steps and ultimately the classifier.
- **Step 3: Select 100 most frequent words**
  Given that the abstract includes many words, word frequency is limited to 100 most frequent words to avoid over-fitting during the classification process.
- **Step 4: Calculate inverse document frequency**
  Inverse document frequency is calculated to identify which words are more informative than others. The calculation is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. The result of this step is a frequency matrix A which is a sparse matrix that is associated with each of the documents and their document frequency for each of those words.
  The inverse document frequency is calculated using Equation 1.

  $$idf(t, D) = \log \frac{N}{|d \in D : t \in d|} = \log \frac{N}{n_t} \tag{1}$$

  where $idf(t, D)$ is a function that calculates frequency of term t in document d. The log factor is to avoid excessive weight to frequent terms.
- **Step 5: Calculate singular value decomposition (SVD)**
  Singular value decomposition of the frequency matrix A is calculated for concept extraction. The main goal of this step is to identify the collection of terms that are more informative when they are paired together than the individual words themselves.
  SVD of matrix A is the product of multiplying an $m \times n$ column orthogonal matrix U with an $n \times n$ diagonal matrix S and an $n \times n$ orthogonal matrix V such that $A = USV^\top$ where m denotes the number of rows and n denotes the number of columns.

C. Phase 3: Classification model

- **Step 1: Train boosted tree (BT) classifier**
  The basic concept of boosted trees is that weak learners are combined as such to create a strong learner. Very simple trees are created before making the final classification. The final classification is provided from the classification of the simple trees as a whole with a learning rate of 0.1. We consider a binary classification, where we classify patents as relevant or irrelevant.

- **Step 2: Automatic classification**
  The algorithm is used to classify new documents by following the steps starting from phase two to phase three.
- **Step 3: Model validation**
  A hold-out is performed to validate the model. The dataset is randomly divided into three subsets which are the training set, validation set, and the test set. Over-fitting is examined by evaluating whether the model fits the training set better than test set. Lift chart and gains chart are also used to validate the model performance.
- **Step 4: Calculate accuracy and performance**
  We will use Equations 2, 3, and 4 to calculate accuracy (acc), precision (p) and recall (r) to determine the model performance. TP denotes true positives which are the correct classification into the relevant class. TN denotes true negatives which are the correct classification into the irrelevant class. FN denotes the incorrect classification into the relevant class, and FP denotes the incorrect classification into the irrelevant class.

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$p = \frac{TP}{TP + FP} \tag{3}$$

$$r = \frac{TP}{TP + FN} \tag{4}$$

IV. Experimental Results and Evaluation

A. Results of Phase 1

We have considered patents related to the development of a service matching system. The system's goal is to match customers to the service providers based on individual case evaluation. The customers in our case are the elders, and the service providers in our case are the health care and the elders' service centers. We have searched Google Patent Search and FreePatentsOnline databases manually for patents based on keywords related to our system. Once we obtained a list of 500 patent data, we calculated the keyword frequencies and created a patent keyword dictionary to query new related patents. The output of this phase was a patent database with a total of 7530 patents.

B. Results of Phase 2

Stop words such as 'a', 'about', 'after', 'by', 'but', 'etc' were removed. The abstract was chosen as the variable of interest. To avoid over-fitting, only the 100 most frequent words where selected. The frequency word count in Table I illustrates some top words, together with the count each word appears in the whole dataset and the total number of files in which each word appears (regardless of the word repetition).

Inverse document frequency matrix was obtained and the singular value decomposition was calculated. Figure 1 demonstrates a graph with singular values. The first three concepts where chosen given that they result in the highest gain, and the remaining concepts were disregarded.

Table I
EXAMPLE OF WORD FREQUENCY COUNTS.

Word	Count	Files
system	8172	3515
provide	7137	3705
health	6721	1775
care	6426	1900
data	6230	1885
include	5539	3402
patient	5167	1273

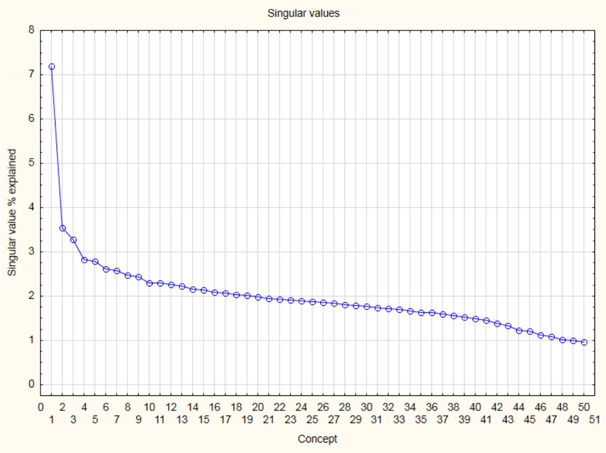

Figure 1. Singular values.

Table II
CLASSIFICATION MATRIX ON TRAINING SET.

Observed	Predicted Relevant	Predicted Irrelevant
Relevant	TP = 169	FP = 32
Irrelevant	FN = 89	TN = 710

The output of this phase is document frequency matrix with the three concepts that we obtained after calculating the singular value decomposition.

C. Results of Phase 3

For the training data, we used the set of document frequency matrix with the results. To measure the strength of our model, we have approached this by the accuracy, precision, and recall. The classification matrix in Table II demonstrates the result of classifying a total of 1000 training set. The model scored accuracy rate of 87.9%, precision rate of 84.07%, and recall rate of 65.50%. The accuracy rate indicates that the model fits well with the data.

After training the classifier, it was found that the optimal number of trees is 189, with a maximum tree size of 3, and that is due to the patents that we want to classify into one of the two classes, that is either relevant or irrelevant. Figure 2 demonstrates that the prediction error for the training data decreased as more trees were added to the model.

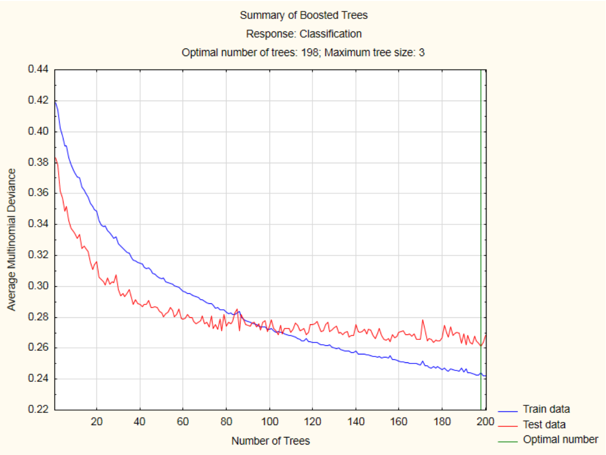

Figure 2. Summary of boosted trees.

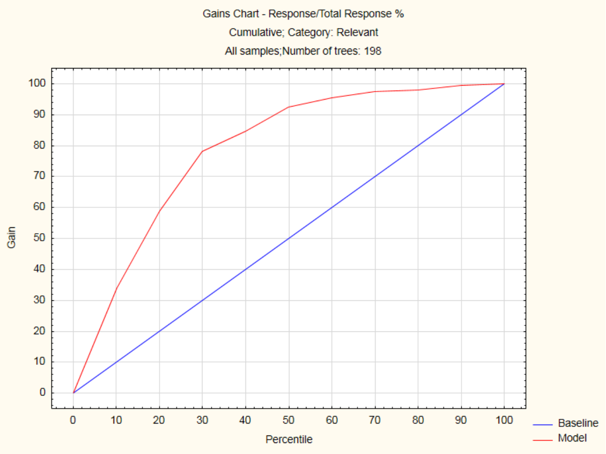

Figure 3. Gains chart for relevant class.

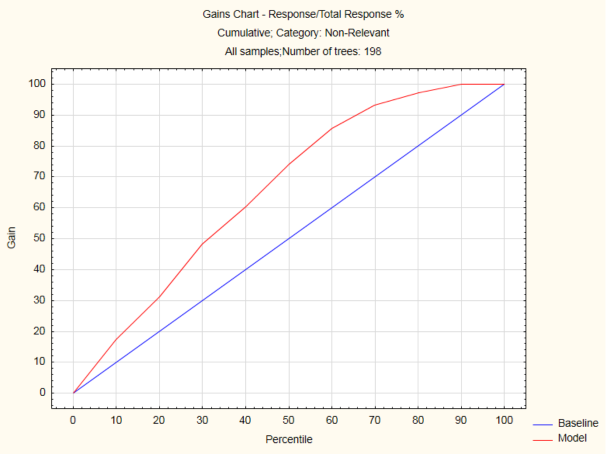

Figure 4. Gains chart for irrelevant class.

To measure the effectiveness of a classification model, gains charts for both relevant and irrelevant classes are illustrated to demonstrate the ratio of the accurate predictions to the total number of patents in that class. Figure 3 demonstrates the cumulative gains chart between the percentile of the total population and respondent gain value for the relevant class. By observing the charts, we can notice that there is less gain for class irrelevant in Figure 4 than the earlier class. Nevertheless, we can summarize relative information from the charts such as that for the first 10% of the percentile, we obtain around 40% gain, and for the next 20% we obtain 68% gain.

From the results, we can conclude that our proposed three-phase model can offer promising outcomes, and thus is a valid approach to conduct automatic patent classification tasks.

V. Conclusion

An affective automatic patent classification system allow companies to correctly identify which patent to invest in and will generate maximum profit. Basic methods such as Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) algorithms were used extensively in previous studies. We have evaluated a three-phase model consisting of patent query (phase 1), text processing (phase 2), and a classification model (phase 3). The main system relies on the output of phase 2 which is the document frequency matrix. The main classification model was Boosted Tree (BT) classifier which scored an accuracy of 87.9% as a result of phase 3. A future direction would be to evaluate the current system by applying alternative classification algorithms such as SVM in phase 3 while following the same steps in the phases 1 and 2.

References

[1] A. J. C. Trappey, C. V. Trappey, C.-Y. Wu, and C.-W. Lin, "A patent quality analysis for innovative technology and product development," *Advanced Engineering Informatics*, vol. 26, no. 1, pp. 26–34, 2012.

[2] J.-L. Wu, P.-C. Chang, C.-C. Tsao, and C.-Y. Fan, "A patent quality analysis and classification system using self-organizing maps with support vector machine," *Applied Soft Computing*, vol. 41, pp. 305–316, 2016.

[3] T. Abdunabi and O. Basir, "Building diverse and optimized ensembles of gradient boosted trees for high-dimensional data," in *Proceedings of the 2014 3rd IEEE International Conference on Cloud Computing and Intelligence Systems*. IEEE, 2014, pp. 351–356.

[4] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *Proceedings of the 1996 International Conference on Machine Learning*, 1996, pp. 148–156.

[5] S. T. Monteiro and R. J. Murphy, "Embedded feature selection of hyperspectral bands with boosted decision trees," in *Proceedings of the 2011 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2011, pp. 2361–2364.

[6] C.-H. Wu, Y. Ken, and T. Huang, "Patent classification system using a new hybrid genetic algorithm support vector machine," *Applied Soft Computing*, vol. 10, no. 4, pp. 1164–1177, 2010.

[7] N. V. Alisova, "Biomedical engineering in international patent classification," *Biomedical Engineering*, vol. 47, no. 3, p. 164, 2013.

[8] M.-L. Antonie and O. R. Zaiane, "Text document categorization by term association," in *Proceedings of the 2002 IEEE International Conference on Data Mining*. IEEE, 2002, pp. 19–26.

[9] C. J. Fall, A. Törcsvári, P. Fiévet, and G. Karetka, "Automated categorization of German-language patent documents," *Expert Systems with Applications*, vol. 26, no. 2, pp. 269–277, 2004.

[10] Y.-L. Chen and Y.-C. Chang, "A three-phase method for patent classification," *Information Processing & Management*, vol. 48, no. 6, pp. 1017–1030, 2012.

[11] R. Eito-Brun, "Knowledge dissemination patterns in the information retrieval industry: A case study for automatic classification techniques," *World Patent Information*, vol. 39, pp. 50–57, 2014.