A Signature Approach to Patent Classification

Dilesha Seneviratne[1], Shlomo Geva[1], Guido Zuccon[1], Gabriela Ferraro[2], Timothy Chappell[1], and Magali Meireles[3]

[1] Queensland University of Technology (QUT), Brisbane, Australia,
d.dwmhakmanawalawwe@hdr.qut.edu.au,{s.geva,g.zuccon,timothy.chappell}@.qut.edu.au
[2] NICTA and Australain National University, Australia,
gabriela.ferraro@nicta.com.au
[3] Pontifical Catholic University of Minas Gerais (PUC Minas), Brazil,
magali@pucminas.br

Abstract. We propose a document signature approach to patent classification. Automatic patent classification is a challenging task because of the fast growing number of patent applications filed every year and the complexity, size and nested hierarchical structure of patent taxonomies. In our proposal, the classification of a target patent is achieved through a k-nearest neighbour search using Hamming distance on signatures generated from patents; the classification labels of the retrieved patents are weighted and combined to produce a patent classification code for the target patent. The use of this method is motivated by the fact that, intuitively, document signatures are more efficient than previous approaches for this task that considered the training of classifiers on the whole vocabulary feature set. Our empirical experiments also demonstrate that the combination of document signatures and k-nearest neighbours search improves classification effectiveness, provided that enough data is used to generate signatures.

1 Introduction

Patents are legal documents issued by governments for giving rights of exclusivity and protection to inventors. Patents play a significant role in helping inventors and organisations to protect their intellectual property. The number of patent applications filed every year is increasing rapidly. For example, in 2012 the World Intellectual Property Organization (WIPO) reported an increase of 9.2% from previous years. Patents are organised in a classification system, called International Patent Classification (IPC), which provides for a hierarchical taxonomy where patents are classified according to the areas of technology to which they pertain. IPC contains about 120 classes and about 630 subclasses. This taxonomy is complex, large and nested (hierarchical), adding to the complexity of the patent classification task.

Given the increasing rate at which patents are filed, the current practice of manually classifying patents is unsustainable due to the time and resource burden it presents [11]. Automated classification systems have therefore emerged; see for example Chakrabarti et al. (multi level Bayesian classifiers) [5,4], Tikk et al. (hierarchical classifiers) [15], Larkey (k-nearest neighbour) [13], Cai and Hofmann (hierarchical classifiers based on SVM) [3], and Chen and Chang (three phase classification) [6]. Of interest to this paper is the work of Fall et al. [8] who have evaluated a number of machine learning classifiers, including support vector

machines (SVM), naive Bayes (NB), and k-nearest neighbour (kNN) classifiers, using bag of words as feature set. Their results suggest that SVM and NB have similar effectiveness when considering the highest level of the IPC hierarchy (class level), while kNN had lower effectiveness. When considering the lower level classification (subclass level), instead, SVM was found to outperform the other classification methods. We shall use the methods explored by this work as a benchmark for our document signature approach.

However, in previous work, little attention has been paid to the *efficiency* of automated methods for patent classification, with improvements in classification effectiveness taking the lion's share of the research efforts. The use of classification techniques such as support vector machines (SVM), however, does not scale to the increasing amount of patents being filed every year. In this paper we address this concern by examining an approach to patent search that is well-known for its efficiency: signature search [9].

Signatures are lengthy bit strings of words that are often created using an hash function. Signatures are used to quickly identify potentially relevant documents. We exploit signatures for patent classification by performing a signature search for patent signatures that are similar to a target patent that is provided as a query for classification. The patents in the k-nearest neighbourhood of the (query) target patent are considered to determine the target's classification code; this is obtained by weighting the classification codes of the patents in the neighbourhood. This approach guarantees extreme efficiency, specifically because of its capacity to scale to very large collections given that indexing is linear with the size of the collection (like inverted file search engines) and searching time increases linearly at a lower rate than the increase in collection size [1].

2 Patent classification with Signatures

Early approaches used to generate document signatures are based on the bitwise OR composition of binary signatures associated with terms in documents [9,7]. Further refinements of the signature generation process have been proposed. A recent approach, called TopSig [10,7], uses random indexing for compressing the standard term-document matrix, followed by aggressive numerical precision reduction to maintain only the sign bits of the projected term-document matrix. This approach has been shown to be superior to standard signature approaches: we thus rely on the TopSig method to generate patent signatures.

Patent signatures are generated from snippets of text extracted from the patents. Specifically, in the experiments of Section 3, we consider the effectiveness of signatures generated either from patent titles, from patent abstracts, from claim texts, or from the first 300 words[4] of patent text.

To perform automated patent classification with signatures, we first construct a signature for the patent requiring classification (target patent). The document signatures are formed through the successive summing of pseudo-randomly generated term vectors created from the patents' text. The resulting

[4] Previous work has also used the first 300 words extracted from each patent: this setting has in fact shown strong promise [8].

document vector is then flattened into a binary signature which functions as a locality-sensitive hash (LSH). We then use this signature to query a collection of patent signatures derived from a training set (in the same manner as the query signature), where patents are labelled with their correct IPC code. This process results in a ranking of patents, ordered in decreasing similarity to the target patent. Similarity in the signature space is measured according to the Hamming distance, i.e., the number of bits in which the two signatures differ.

To determine the first level of classification (section), the k-nearest neighbour classification algorithm is employed. The k-nearest neighbourhood to the target patent signature is formed by selecting the top-k patents from the ranking. A classification label is then produced for the target patent by a majority vote of its neighbours, with the label being selected from the class most common amongst its k nearest neighbours measured by the distance function $w(p) = \frac{1}{\sqrt{rank(p)}}$, where $rank(p)$ is the rank at which patent p has been retrieved in answer to the query formed by the target patent's signature. This procedure is akin to a simple voting process, where each training patent in the k-neighbourhood votes for its label, votes for the same label cast by different patents are accumulated and modulated by a weight $w(p)$ inversely proportional to the rank of the voting patent in the neighbourhood ranking.

To determine further levels of classification (class, subclass, group), the voting process is iterated but considering only subsets of patents in the k-neighbourhood that share the same higher level label as that assigned to the target patent. For example, when determining the class label for a target patent, the training patents that are considered are only those in the k-neighbourhood that share the section label assigned by our method to the target patent are considered.

3 Experiment Settings

3.1 Dataset

To evaluate the proposed approach to patent classification based on signatures we use the WIPO-alpha collection, a standard collection for patent classification used also by previous work. The WIPO-alpha collection (WIPO in short) consists of over 75,000 patent applications that have been submitted to WIPO under the Patent Cooperation Treaty (PCT). The collection is split into train and test sets which consist of 46,324 and 28,926 patents respectively.

3.2 Evaluation measures

In line with previous work [8,15], to evaluate the effectiveness of the automated patent classification approaches, we analyse the number of correct guesses made by the classifiers when compared with the ground truth (precision). The (micro-average) precision is computed according to three settings: (1) the *top prediction* made by each classifier, where the classifier prediction with highest score is compared with the classification label recorded in the ground truth; (2) the top *three guesses* made by each classifier, where the classifier predictions with the three highest scores are compared with the classification label recorded in the ground truth, and success is recorded if one of these prediction does match with the ground truth; and (3) the *All Categories* method, where the top prediction of

the classifier is compared with all the categories recorded in the ground truth, in case one match is found, the classification is deemed successful. Note that generally patents are assigned to multiple classification codes.

3.3 Approaches and Settings

For the signature based approach, we first preprocessed the patents by removing characters that are not alphabetic; we also removed stop words and applied the Porter stemmer. To create signatures, we set the signature width to 4096 bits. For the kNN classifier, we set k to 30 following the benchmark approach [8].

The effectiveness of the signature based approach is compared to that achieved by the classifiers investigated by Fall et al. [8] because they also used the WIPO-alpha collection and considered all classification codes (rather than limiting to particular sections of the IPC hierarchy). Moreover, that work explored the effectiveness of the kNN algorithm and thus we can directly compare the benefits of using signatures over bag of words.

4 Results

4.1 Effectiveness

Table 1 reports the results obtained by benchmarks and proposed approach at the IPC class level classification, while the results obtained at the IPC subclass level evaluation are reported in Table 2. The results were obtained by considering different text snippets to create representations of patents, either based on bag of words (for the benchmark methods) or document signatures (for the proposed method); these are: title of patent, abstract of patent(including titles, inventors, applicants) first 300 words of the patent text (titles, inventors, applicants, abstracts, and descriptions).

The results highlight that all classification methods are less effective when classifying at lower granularity levels (subclass) than at higher granularity (class). For the kNN method with document signatures, this is because classifications for low granularity levels are affected by those obtained at higher granularities: thus, if an error is made at class level, the error is propagated to subclass and group level. For the benchmark methods, loss in effectiveness is instead generally due to less training data being available for subclass level classification than at class level classification.

When signature and benchmark methods are compared, we observe that the signature method is always more effective than its direct counterpart in the bag-of-words space, i.e., the kNN classifier of Fall et al. [8]. When other benchmark

Table 1. Classification results at IPC class level.

Indexing field	Evaluation measures	NB Fall et al. [8]	k-NN Fall et al. [8]	SVM Fall et al. [8]	k-NN Proposed method
Title	Top-prediction	45%	33 %	Not reported	**40%**
First 300 words	Top-prediction	55%	51%	55%	**56%**
Title	Three-guesses	66 %	52%	Not reported	**63%**
First 300 words	Three-guesses	79%	77%	73%	**81%**
Title	All-categories	52 %	38%	Not reported	**46%**
First 300 words	All-categories	63%	58%	62%	**63%**

Table 2. Classification results at IPC subclass level.

Indexing field	Evaluation measures	NB Fall et al. [8]	k-NN Fall et al. [8]	SVM Fall et al. [8]	k-NN Proposed method
Abstract	Top-prediction	28%	26%	34%	**32%**
First 300 words	Top-prediction	33%	39%	41%	**42%**
Abstract	Three-guesses	47%	45%	52%	**53%**
First 300 words	Three-guesses	53%	62%	59%	**67%**
Abstract	All-categories	35%	32%	41%	**38%**
First 300 words	All-categories	41%	46 %	48%	**50%**

methods are considered instead, SVM and NB are found to be more effective that the signature based kNN when title snippets are used to generate patent representations. However, if longer snippets are used, as is the case when using the first 300 words of a patent, then the classification precision increases (for both proposed and benchmark methods); more importantly, the effectiveness of the proposed approach reaches (and can even outperform) that of benchmark methods. This suggests that signature approaches are comparable (or sometimes superior) to bag-of-words approaches in terms of classification effectiveness if enough evidence is used to produce the patent representations. This is more evident when considering top three predictions (Tables 1 and 2). Moreover, the results suggest that if more than one classification could be assigned to a patent, then effectiveness increases on average of approximately 20%. Finally, when we analysed the effectiveness of the classifiers with respect to each IPC section (not reported for brevity), we found that the section where all classifiers delivered the lowest precision was section G (Physics); previous work had reported similar findings [8].

4.2 Efficiency and Scalability

The use of document signatures as an alternative to bag-of-word features for patent classification was motivated by the fact that document signatures provide significant advantages in terms of search times and scalability. Table 3 reports the time required to index the WIPO-alpha collection and that required to search in order to perform classification[5]. These results highlight that signatures allow searching patent collections within milliseconds and that the increase in the amount of text that is represented by a signature does not result in a large increment in time required to search for similar signatures. To study whether the document signature approach is scalable to larger patent collections, we replicated the classification experiments (for abstract only) for the USPTO collection, a dataset of more than 1.4 million patents (thus three orders of magnitude larger than WIPO). The runtime results (Table 3) highlight the scalability of the signature approach, since querying time increased of only one order of magnitude while the collection increased of three orders of magnitude.

[5] No publicly available implementation of Fall et al.'s methods was available and our re-implementation did not lead to effectiveness comparable to the reported one. We were therefore unable to obtain efficiency figures for the benchmark methods. Similarly, we were unable to test for significant differences.

Table 3. Time required to index and search a patent collection.

Collection	Field	Indexing time	Searching time (Avg per query)
WIPO	Abstracts	4.43 s	2.8×10^{-2}s
Train-46,324	Title	1.40s	1.6×10^{-4}s
Test- 28,926	First 300 words	9.14s	6.9×10^{-2}s
USPTO (2006-2013)			
Train-1,358,908	Abstracts	68.96s	4.5×10^{-1}s
Test-51,324			

5 Conclusions

In this paper we have investigated the use of document signatures for patent classification. Our initial empirical experiments have provided a number of interesting insights on the use of document signatures for this classification task and have opened avenues for future work. The results continued that the signature approach provides an efficient and scalable solution for this problem, and this is highly comparable in terms of effectiveness with other approaches to patent classification like SVM (which in turn do not scale to large patent collections). Moreover, our initial experiments have shown that the selection of which part of the patent is used to generate signatures is fundamental for the effectiveness of the classifiers. More research is however needed to understand what is the best content/part of the patent that should be used to generate signatures: titles, abstracts and the first 300 words gave correspondingly different results.

References

1. T. Chappell, S. Geva, and G. Zuccon. Approximate Nearest-Neighbour Search with Inverted Signature Slice Lists In *Advances in Information Retrieval*, pages 147–158, 2015.
2. K. Benzineb and J. Guyot. Automated Patent Classification. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 239–261, 2011.
3. L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of CIKM '04*, pages 78–87, 2004.
4. S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. In *VLDB*, volume 97, pages 446–455, 1997.
5. S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7(3):163–178, 1998.
6. Y.-L. Chen and Y.-C. Chang. A three-phase method for patent classification. *Information Processing & Management*, 48(6):1017 – 1030, 2012.
7. C. M. De Vries and S. Geva. Pairwise similarity of topsig document signatures. In *Proceedings of ADCS '12*, pages 128–134, 2012.
8. C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. Automated categorization in the international patent classification. *SIGIR Forum*, 37(1):10–25, Apr. 2003.
9. C. Faloutsos. Signature-based text retrieval methods: A survey. *Data Eng.*, 13(1):25–32, 1990.
10. S. Geva and C. M. De Vries. Topsig: Topology preserving document signatures. In *Proceedings of CIKM '11*, pages 333–338, 2011.
11. J.-H. Kim and K.-S. Choi. Patent document categorization based on semantic structural information. *Information Processing & Management*, 43(5):1200 – 1215, 2007. Patent Processing.
12. M. Krier and F. Zaccà. Automatic categorisation applications at the european patent office. *World Patent Information*, 24(3):187 – 196, 2002.
13. L. S. Larkey. A patent search and classification system. In *Proceedings of DL '99*, pages 179–187, 1999.
14. H. Smith. Automation of patent classification. *World Patent Information*, 24(4):269 – 271, 2002.
15. D. Tikk. A hierarchical online classifier for patent categorization. pages 244–267, 2007.