A Semantic Approach to Patent Mining for Relating IPC to a Research Paper Abstract

Md. Hanif Seddiqui[1], Yohei Seki[1], Masaki Aono[1]

[1]Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, Japan
hanif@kde.ics.tut.ac.jp, {seki, aono}@ics.tut.ac.jp

Abstract. *If the business entities can identify research gaps and predict research trends prior to their investments on the research of their interests, they can profitably and strategically invest their research fund to acquire more and more patents. The primary step of identifying research gaps and predicting research trends is to relate International Patent Classification (IPC) to a research paper abstract, because it is believed that there is a close relationship between patents and state-of-the-art research papers. Naively relating IPC to a scientific paper abstract using huge amount of patent documents is a formidable task due to the massive amount of the patent documents and due to the varieties of field specific technical terminologies. Our research proposes an efficient semantic approach to patent mining that retrieves IPC related to a research paper abstract by combining the following data and technologies, i.e,. an ontology of IPC, ontology alignment techniques, and prior knowledge of the relation between IPC and terminologies inside patent documents. Our system retrieves probable IPCs related to an abstract from prior knowledge. Then the neighboring concepts of the probable IPCs are retrieved from the ontology for acquiring terminological and semantic similarities between the text used in IPC and an abstract. Our system has a salient feature of efficient computation to relate IPC to scientific paper abstract. Preliminary experiments show that our system outperforms a baseline system, which "naively" relates IPC to a research paper abstract.*

Keywords: Ontology, Semantic Technology, Patent Mining, international Patent Classification

1 Introduction

The immerse growth of patent documents necessitates powerful algorithms and tools that can automatically perform mining of patent like patent categorization to relate to a research paper abstract. The mining of patent documents, specially relating patent classification to a research paper abstract, identifying research gap and predicting research trends to the potential inventors, researchers, development units and even to the patent issuing authorities prior to their intensive attention on the research.

The patent offices organize patent applications into very large topic taxonomies. The most important among them is International Patent Classification (IPC). The World Intellectual Property Organization (WIPO) maintains IPC within an ontology in XML format having concepts taxonomies and cross references as concept relations. The IPC taxonomy consists of about 80,000 categories that cover the whole range of industrial technologies. There are eight sections named *A* through *H* at the highest level of the hierarchy, then 128 classes, 648 subclasses, about 7200 main groups and 72000 subgroups at the lower levels (See Fig. 1). The subgroups are even classified into different levels. The top four levels are usually the target of automated patent categories excepts there are very few which is described in Section 2.

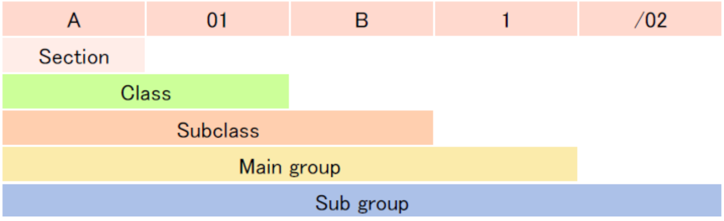

Fig. 1: *A* is a section for 'Human Necessities', *A01* is class representing 'Agriculture; Forestry; Hunting; Fishing; etc.', *A01B* is subclass which consists of 'Soil working in agriculture or forestry etc.', *A01B 1/00* is a main group representing 'Hand Tools', while *A01B 1/02* is a subgroup for 'Spades; Shovels'.

The collection of patent documents is also quite large. We have eight years English patent documents from 1993 through 2000, which includes about one million of patent documents. An average patent document contains more than 3000 words. Moreover, many vague and general terminologies are often used to avoid narrowing the scope of the invention [8]. Combining a general terms are often have a special meaning that also has to be captured. Patent document contains even acronyms and much new terminology [13].

Due to these factors, it is difficult to discern required information manually, thus, patent analysis has long been considered as useful in product innovative process to identify research gap where complementary technology can be licensed in or to identify the related researches being licensed so far. To achieve the goal of patent mining, machine learning and text mining techniques are widely used in patent analysis. As patent documents are huge in number, it is obviously not worthy task to consider every of one million patent documents while patent mining. Moreover, indexing of terminologies is not sufficient in patent mining system as tendency of using vague and more general terminologies. The overriding philosophy of a classification scheme is to identify a single point for each document or abstract within the *universe of knowledge*. Consequently, when a document discloses multiple concepts, rules of precedence have to be applied in order to determine the final classification of sufficient depth [1].

To overcome the problems of linear system of automatic categorization of patent classification, our system has a unique part which uses an ontology of IPC as a universe of knowledge which is instantiated by the terminologies of huge patent documents to enrich the knowledge base of IPC ontology. Then our semantic approach only consider a small part of huge IPC documents and a small part of IPC taxonomy with the help of keyword-IPC knowledge and with our developed ontology alignment algorithm to focus only a specific part of large taxonomy taking advantages of *locality of references*. Eventually, our system can produce more relevant IPC in sufficient depth for a research paper abstract with the help of ontology defined in semantic technology and utilizing the techniques of ontology alignment. It is capable of generating significantly better categorization results within short elapsed time.

The rest of the paper is organized as follows: in Section 2, related works are described. Section 3 focuses our patent mining system for patent categorization, while Section 4 contains the experimental results. Concluded remark and future works are described in section 5.

2 Related Works

From the late 1990s, machine learning techniques of text categorization [22] received increasing attention in automatic categorization of patent classification. The categorization of the patent classification scheme can be performed in two ways: an algorithm can either flatten the taxonomy and consider it a system of independent categories or can incorporate the hierarchy in the categorization algorithm. Early patent categorizers chose the former solution, but these were outperformed by real hierarchical classifiers. The first hierarchical classifier was developed by Chakrabarti et al [3, 4] using Bayesian hierarchical classification system applying the Fisher's discriminant. The Fisher’s discriminant is a well-known technique from statistical pattern recognition. It is used to distinguish feature terms from noise terms efficiently. They tested the approach on a small-scale subtree of a patent classification consisting of 12 subclasses organized in three levels. Here they found that by using the already-known classifications of cited patents in the application, the effectiveness of the categorization could be much improved [5]. Larkey [17, 18] has created a tool for attributing US patent codes based on a k-Nearest Neighbor (k-NN) approach. The inclusion of phrases (multi-word terms) during indexing is reported to have increased the system’s precision for patent searching but not for categorization [17], though the overall system precision is not specified. Kohonen et al [14] developed a self-organizing map based PC system. Their baseline solution achieved a precision of 60.6% when classifying patents into 21 categories. This could be raised to 64% when different feature selection techniques have been applied. A comprehensive set of patent categorization tests is reported in [16]. These authors organized a competitive evaluation of various academic and commercial categorizers, but have not disclosed detailed results. The participant with the best results has published his findings separately [15]. They implemented a variant of the Balanced Winnow, an online classifier with a multiplicative weight updating schema. Categorization is performed at the level of 44 or 549 categories specific to the internal administration of the European Patent Office, with around 78% and 68% precision, respectively, when measured with a customized success criterion. The above listed approaches are difficult to compare given the lack of a benchmark patent application collection and a standard patent taxonomy. This lack has been at least partly alleviated with the disclosure of the WIPO document collections. First, the WIPO-alpha English collection was published in 2002 [10], and shortly after the WIPO-de German patent application corpus became publicly available [9]. The creators of the WIPO-alpha collection [8] performed a comparative study with four state-of-the-art classifiers (Naive Bayes, NB; Support Vector Machine, SVM; k-NN and a variant of Winnow) and evaluated them by means of performance measures customized to typical PC scenarios. The authors found that at the class level NB and SVM were the best (55%), while at the subclass level SVM outperformed other methods (41%). Since then, several works reported results on WIPO-alpha. Unfortunately, most authors scaled down the problem by working only on a subset of the whole corpus. Hofmann et al [12] experimented on the D section (Textile) with 160 leaf level categories and obtained 71.9% accuracy. Rousu et al [20] evaluated their SVM-like maximum margin Markov network approach also on the D section of the hierarchy, and achieved 76.7% averaged overall F-measure value. Cai & Hofman [2] tested their hierarchical SVM-like categorization engine on each section of WIPO-alpha, and obtained 32.4–42.9% accuracy at the maingroup level. Godbole & Sarawagi [11] presented another SVM variant that has been evaluated on the entire hierarchy and specifically on the F subtree (Mechanical engineering, lighting, heating, weapons, blasting) of the corpus. They achieved 44.1% and 68.8% accuracy, respectively.

A patent application oriented knowledge management system has been developed by Trappey et al [23], which incorporates patent organization, classification and search methodology based on back-propagation neural network (BPNN) technology. This approach focuses on the improvement of the patent document management system in terms of both usability and accuracy. The authors compared their method with a statistical and a Bayesian model and found some improvement in accuracy when tested again a small-scale two-level subset of the WIPO-alpha collection (a part of B25; Power hand tools) with 9 leaf level categories. The paper put special emphasis on the extraction of key phrases from the document set, which are then used as inputs of the BPNN classifier. Other hierarchical categorization algorithms such as in [6], [21], or [7], have not been evaluated on patent categorization benchmarks.

3 Our System

Our system uses ontology in the form of taxonomy from semantic technology. The ontology of the semantic essence improves the performance and results of the automatic categorization of the patent classification. Our system includes two major steps for the whole process: preprocessing and the main processing.

3.1 Preprocessing

Our system contains two preprocessing units. One unit is for the creation of hierarchy of International Patent Classification (IPC) from the IPC data available in XML format at the WIPO site[1] (See Fig. 1a).

We also develop the efficient feature vector (See Fig. 1b). Almost one million English patent documents are available in a dataset from the year 1993 through 2000. Our text classifier represents a document as a set of features, $d=\{f_1, f_2, f_3, ...f_m\}$, where m denotes the number of active features that occur in the documents and every patent document is associated with primary IPC. Feature, typically, represents a word or a word-phrase (sequence of words). The relevance of feature f in a specific category of patent classification, c is given by the weight $w(f, c)$, which is measured by TF-ICF model depending on the number of times f occurs in the category and the inverse category frequency as follows:

$$w(f,c) = TF(f,c) * \log\left(\frac{N}{|f \in c|}\right),$$

where N, denotes the total number of categories and the denominator of the logarithm denotes the number of categories a feature, f belongs to.

Therefore, a primary prior knowledge is represented by feature vector where each feature is associated to categories with their TF-ICF weights. However, the vector may contain general features, which will lead the model to misclassification. In order to solve the problem, we consider *effectiveness* of the features, modification of the weight, and the method of Littlestone's Positive Winnow.

If a feature is available into more than one document in a specific category, and not available in other documents of different categories is considered as the most effective feature. We remove all features which belongs to more than two categories or available only one document in one category.

ICF plays an important role to determine generality of features. The more general feature has lower the value of ICF. Therefore, we use ICF^2 to differentiate general and effective features as it will convert high value to higher compared to the other. As a result, the modified weight measure becomes

$$w(f,c) = TF(f,c) * \left(\log\left(\frac{N}{|f \in c|}\right)\right)^2$$

If a classifiable abstract contains features, $a=\{f_{a1}, f_{a2} \ldots f_{an}\}$, then the classifiers can evaluate the similarity between the classifiable abstract and categories by calculating-

$$\Phi_c(a) = \sum_{f \in a} w(f,c) * TF(f,a),$$

where $TF(f, a)$ is the frequency of feature f in the abstract a.

The weighted factors of features are modified by mistake driven online learning model first proposed in [19]. Mistake driven algorithms have typically three parameters: a threshold θ, a promotion parameter α, and a demotion parameter β.

---

[1] International Patent Classification (IPC), http://www.wipo.int/classifications/ipc/en/UT

After initializing the category weights directly from the huge patent documents, Littlestone's *Positive Winnow* assigns a document to a category *iff*: $\sum_{f\in d} w(f,d).w(f,c) > \theta$. The algorithm performs multiplicative weight updating on active features with $\alpha>1$ and $0<\beta<1$. Positive Winnow updates the category weights in the following two cases of mistakes:

1. True label is not found: If the algorithm guesses 0 and the true label is 1 then all active weights are promoted by multiplying them withα.
2. Misclassification: If the algorithm guesses 1 but the true label is 0 then all active weights are demoted by multiplying them withβ.

In both cases the other, non-active weights remain unchanged.

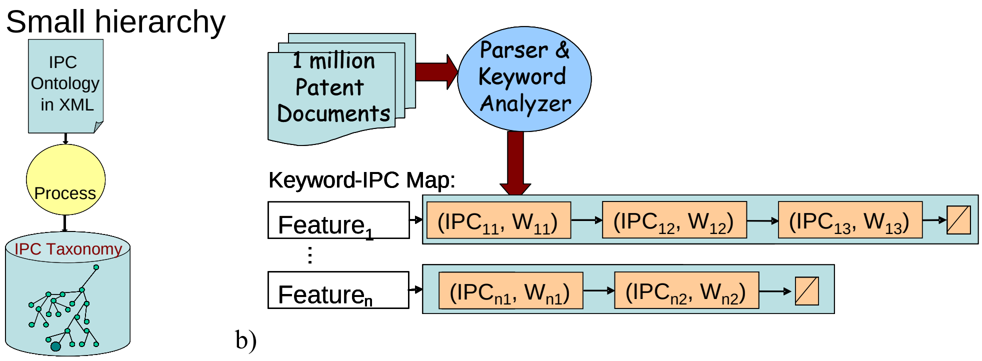
Small hierarchy
IPC Ontology in XML
Process
IPC Taxonomy
1 million Patent Documents
Parser & Keyword Analyzer
Keyword-IPC Map:
Feature_1
$(\text{IPC}_{11}, \text{W}_{11})$
$(\text{IPC}_{12}, \text{W}_{12})$
$(\text{IPC}_{13}, \text{W}_{13})$
Feature_n
$(\text{IPC}_{n1}, \text{W}_{n1})$
$(\text{IPC}_{n2}, \text{W}_{n2})$
b)

Fig. 1. a) creation of IPC taxonomy, b) creation of persistent Feature-IPC knowledge

Preprocessing phase results hierarchy of IPC or taxonomy of IPC that includes some cross field references and this phase also produced feature-IPC mapping that plays an important role in the system main process.

3.2 Methodologies

The Taxonomy and the feature-IPC mapping are persistent data model in main memory. Our system retrieves features from the research paper abstract by text classifier. We have layered model of feature to identify section, class, subclassFrom and IPC as a whole. Our experiments depicts that sequential identification of section, class, and subclass has positive impact over the results. Our linguistical methods can retrieve correct section, class and subclass most of the time. However, it has limitations retrieving more specific (deeper in the hierarchy) IPCs. Therefore the IPCs by linguistic methods are not considered as final output, rather it is considered as primary probable IPCs. Fig. 2 depicts the overall flow of the overall methodologies.

The IPC taxonomy is the main strength of our system. To obtain more specific and accurate IPCs, our system considers the probable IPC as an anchor point of further calculation of the similarities.

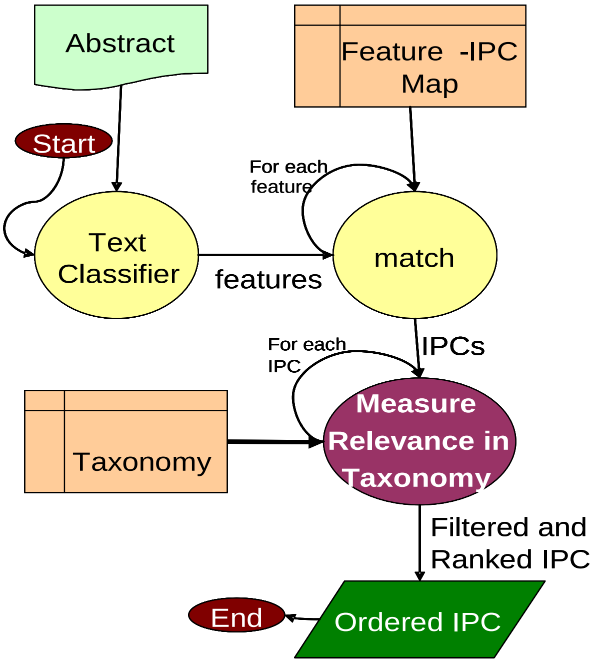

Fig.2 The overall block diagram of our patent mining system which produces ranked list of proposed IPCs for a research paper abstract.

Starting from an anchor IPC in the taxonomy, our system traverses towards the ancestors of upto the subclass level, siblings IPCs, the descendants and the referenced IPCs explicitly defined in the taxonomy (See Fig. 3). Then, our system recalculate the similarities between the features of the research paper abstract and the prototype document of each IPC. The IPC above the threshold, θ are selected. Among the selected IPCs, similarity values are propagated downwards only. Then the ranked IPCs are resulted based on their similarity values after propagation.

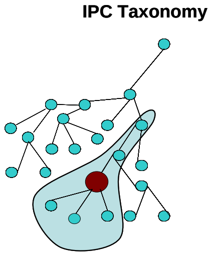

Fig. 3 Traversing neighbors from the anchor (red entity) to see the semantic relatedness

4 Experiments

We applied our system on around 80 thousand research abstracts provided for the English patent mining subtask of NTCIR-7[2] and the average run time is around 0.5 seconds for each abstract.

Table 1. A small part of very large output which shows the strength of IPC taxonomy to remove large number of unrelated IPCs.

Abstract Serial Number	Number of Primary IPC	Number of Ranked IPC using Taxonomy	Time Required
1	7,146	23	531
2	5,904	27	516
3	3,831	19	468
4	4,102	20	454

In terms of precision, our system only uses 100 abstract for this time and we see that it is capable of producing 77.12% precision automatically at subclass level, while considering full taxonomy, it produces precision of 57.23% which is clearly outperforms other systems in the automatic categorization of patent classification.

5 Conclusions and Future Works

In this paper, we describe a system to retrieve related ranked IPC for a resaerch paper abstract by using ontology of semantic technology. Using the semantic technology, our system results more relevant IPC effectively and quickly. We measured similarities between the sets of features from a research paper abstract and a prototype document of a IPC category. The prototype documents are used as prior knowledge towards retrieving probable IPCs. Our system performs well due to the capability of using ontology and due to look at the semantic around an IPC by considering locality of reference. Although our algorithm is still naïve at utilizing the essence of ontology effectively, locality of reference helps to produce better results and to run faster.

Our future target is to enhance the utilization of ontology and to evaluate the results with more correct result-set to measure further precision.

References

1. Adams, S.: Using the International Patent Classification in an online environment, World Patent Information 22(4), 291-300 (2000)
2. Cai, L., Hofmann, T.: Hierarchical document categorization with support vector machines. In Proc. of the 13th ACM Int. Conf. on Information and Knowledge

---

[2] http://research.nii.ac.jp/ntcir/

Management (CIKM'04) (pp. 78–87). Washington D.C.: ACM Press, New York, NY (2004)
3. Chakrabarti, S., Dom, B., Agrawal, R., Raghavan P.: Using taxonomy, discriminants, and signatures for navigating in text databases. In Proc. of 23rd VLDB conference (pp. 446–455), Athens, Greece: Morgan Kaufmann (1997)
4. Chakrabarti, S., Dom, B., Agrawal, R., Raghavan, P.: Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies, VLDB Journal, 7(3), 163–178 (1998)
5. Chakrabarti, S., Dom B., Indyk, P.: Enhanced hypertext categorization using hyperlinks. In Proc. SIGMOD98, ACM International Conference on Management of Data (pp. 307–318), Seattle, WA: ACM Press, New York (1998)
6. Dekel, O., Keshet, J., Singer, Y.: Large margin hierarchical classification. Proc. of the 3rd Int. Conf. on Machine Learning and Cybernetics (ICML'04), (pp. 209–216), Banff, AB, Canada: Morgan Kaufmann (2004)
7. Dumais, S. T., Chen, H.: Hierarchical classification of web content. Proc. of 23rd ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR'00), (pp. 256–263), Athens, Greece: ACM Press, New York (2000).
8. Fall, C. J., Törcsvári, A., Benzineb, K., Karetka, G.: Automated categorization in the international patent classification. ACM SIGIR Forum archive, 37(1), 10–25 (2003).
9. Fall, C. J., Törcsvári, A., Fievét, P., Karetka, G.: Additional readme information for WIPO-de autocategorization data set. http://www.wipo.int/ibis/datasets/wipo-de-readme.html (2003)
10. Fall, C. J., Törcsvári, A., Karetka, G.: Readme information for WIPO-alpha autocategorization training set. http://www.wipo.int/ibis/datasets/wipo-alpha-readme.html (2002)
11. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In Dai, H., Srikant, R., & Zhang, C. (Eds.), Proc. of the 8th *Pacific-Asia Conf. on Knowledge Discovery and Data Mining* (PAKDD'04) (pp. 22–30), Sydney, Australia: Springer-Verlag, Berlin Heidelberg, LNAI 3056 (2004).
12. Hofmann, T., Cai, L., Ciaramita, M.: Learning with taxonomies: Classifying documents and words. In Workshop on Syntax, Semantics, and Statistics (NIPS'03). Whistler, BC, Canada (2003).
13. Kando, N.: What shall we evaluate? Preliminary discussion for the NTCIR Patent IR Challenge based on the brainstorming with the specialized intermediaries in patent searching and patent attorneys, Proc. ACM-SIGIR Workshop on Patent Retrieval (pp. 37–42). Athens, Greece: ACM Press, New York (2000).
14. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., Saarela, A.: Self organization of a massive document collection, IEEE Trans on Neural Networks, 11(3), 574–585 (2000).
15. Koster, C. H. A., Seutter M., Beney, J.: Classifying Patent Applications with Winnow, In Proc. of Benelearn 2001 Conf. (pp. 19–26), Antwerpen, Belgium (2001).
16. Krier, M., Zaccà, F.: Automatic categorization applications at the European Patent Office, World Patent Information, 24, 187–196 (2002).
17. Larkey, L. S.: Some issues in the automatic classification of U.S. patents, In Working Notes for the Workshop on Learning for Text Categorization, 15th Nat. Conf. on Artificial Intelligence (AAAI-98), Madison, WI (1998).
18. Larkey L. S.: A patent search and classification system, In Proc. of DL-99, the 4th ACM Conference on Digital Libraries (pp. 179–187), Berkeley, CA: ACM Press, New York (1999).
19. Littlestone, N.: Learning quickly when irrelevant attributes around: A new linear-threshold algorithm. Machine Learning, 2, 285–318 (1988).

20. Rousu, J., Saunders, C., Szedmak, S., Shawe-Taylor, J.: Learning hierarchical multi-category text classification models, Proc. of the 22nd Int. Conf. on Machine Learning (pp. 745–752), Bonn, Germany: Omnipress (2005).
21. Ruiz, M. E., Srinivasan, P.: Hierarchical text categorization using neural networks. Information Retrieval, 5(1), 87–118, Kluwer Academic Publishers (2002).
22. Sebastiani, F.: Machine learning in automated text categorization, ACM Computing Surveys (CSUR), 34(1), 1-47 (2002).
23. Trappey, A. J. C., Hsu, F.-C., Trappey, C. V., Lin C.-I.: Development of a patent document classification and search platform using a back-propagation network. Expert Systems with Applications, 31(4), 755–765 (2006).