Ontology based Approach to Patent Mining for Relating International Patent Classification (IPC) to a Scientific Abstract

Md. Hanif Seddiqui Yohei Seki Masaki Aono
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, Japan
hanif@kde.ics.tut.ac.jp, seki, aono@ics.tut.ac.jp

November 15, 2008

Abstract

Identifying research gap and predicting research trend is a formidable task in the field of patent mining. The primary step to accomplish the task is to relate International Patent Classification (IPC) to a research paper abstract. Naively relating IPC to a scientific paper abstract is not an easy task due to the generality of terms available in an abstract, the massiveness of the patent documents and the availability of innovative new field specific technical terminologies. Our research proposes an efficient ontology approach to patent mining that retrieves IPC related to a scientific abstract by combining the data and the methodologies used in the field of ontology. The data contains an ontology of IPC and terms to IPC mapping. First, the system uses the extracted terms to retrieve probable IPCs from the terms to IPC mapping. We consider each of the probable IPCs as an anchor point in IPC ontology for further analysis. Our system starts aligning terms available in abstract to the hierarchy of the ontology of IPC to detect correct IPCs and to remove irrelevant one. Our system has a salient feature of efficient computation to relate IPC to scientific paper abstract. The way of using IPC ontology in retrieving related IPC is a novel process.

Keywords: Ontology, ontology alignment, Patent Mining, International Patent Classification (IPC).

1 Introduction

The immerse growth of patent documents necessitates powerful algorithms and tools that can automatically perform patent mining like patent categorization to relate International Patent Classification (IPC) to a scientific abstract, identifying research gap and predicting research trends. The patent mining is becoming important to the potential inventors, researchers, development units and even to the patent issuing authorities. The primary task is the patent categorization of a document like an abstract. In this connection, we have a large IPC taxonomy organized by World Intellectual Property Organization (WIPO) and huge number of classified patent documents. WIPO maintains IPC within an ontology in XML format [1] having concepts taxonomies and relations like cross references. 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.

Figure 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'.

Moreover, we have large collection of preclassified English patent documents of eight years from 1993 through 2000, which includes about one million of patent documents. An average patent document contains more than 3000 words. However, many vague and general terminologies are often used to avoid narrowing the scope of the invention [8]. Patent document contains even acronyms

[1] http://www.wipo.int/classifications/ipc/en/download_area/20080101/xml/ipcr_scheme_20080101.zip

and many new technical terminologies [13], which make patent mining task challenging. Therefore, to achieve the goal of patent mining, machine learning and text mining techniques are widely used for patent analysis in the past. As patent documents are huge in number, it is obviously not worthy task to consider each of one million patent documents while patent mining. Moreover, indexing of terminologies is not sufficient in patent mining system as the 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 of IPCs, rules of precedence have to be applied in order to determine the final classification of sufficient depth [1]. Some effective technique of disambiguation is necessary then.

To overcome the problems of automatic patent classification, our system introduces a new approach. Our system uses ontology of IPC available in the WIPO official website, creates model of taxonomy for IPCs. It also generates mapping between terms available in patent documents to the preclassified IPC. First, our system uses the term to IPC mapping to retrieve probable IPCs related to a given abstract or document. We consider each of the probable IPCs as an anchor point to start off finding further similarity between the abstract terminologies to the description of neighboring IPCs. It refines the probable IPCs taking advantages of the locality of references. Eventually, our system can produce more relevant IPC in sufficient depth for a scientific abstract with the help of ontology and utilizing the techniques of ontology alignment. Theoritically, it is capable of generating significantly better categorization results within short elapsed time.

We organize the rest of the paper as follows: Section 2 contains the description of the related works. Section 3 focuses our patent mining system for patent categorization, while Section 4 contains the experimental results. We conclude about our research and future works in section 5.

2 Related Works

From the late 1990s, machine learning techniques 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 systems 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 [9], and shortly after the WIPO-de German patent application corpus became publicly available [10]. The creators of the WIPO-alpha collection [8] performed a comparative study with four state-of-the-art classifiers (Naive Baye's, 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. [21] 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 and 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 and 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, 7, 22] have not been evaluated on patent categorization benchmarks.

3 Our System

Our system uses ontology of semantic technology in the form of taxonomy. The ontology usually 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

The preprocessing steps contains two independent operations, such as creating IPC taxonomy from IPC ontology available in XML format, creating terms to IPC mapping from huge patent documents by the methods of text mining.

3.1.1 Creating Taxonomy of IPC

Our system creates the taxonomy with some simple relations of International Patent Classification (IPC) from the IPC data available in XML format at the WIPO site. We used DOM XML parser to parse the IPC contents. The XML file for IPC contains entryReference tag for referencing other other IPCs relatively similar, but from different group. We parsed the entryReference tag as a relationship between IPCs. The relationship in the taxonomy of IPC is the unique consideration, which deals many indirect categorization of patent classification.

3.1.2 Creating Mapping between Terms and IPC

We also develop the efficient feature vector. Almost one million preclassified 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, $d = \{f1, f2, f3, ...fm\}$, where m denotes the number of active features that occur in the documents and every patent document is associated with a primary IPC. Feature, typically, represents a word or a word-phrase (sequence of words) having unique meaning together. 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 $TF - ICF$ (TF stands for *"Term Frequency" and ICF stands for* "Inverse Category Frequency") model depending on the number of times f occurs in the category and the inverse category frequency as follows:

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

where N, denotes the total number of categories and the denominator in the logarithm denotes the number of categories a feature, f belongs to. Therefore, a primary term-IPC mapping 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 by modifying the weight, and the method of Littlestone's Positive Winnow [19].

We also measure the effectiveness of a feature. 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 effective feature. We remove all features which belongs to more than two categories or available only one document in one category.

On the other hand, 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

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

The preprocessing is only measured once and kept as a repository. It is reused until the IPC taxonomy is changed by WIPO or any new patent documents come out. After the preprocessing IPC taxonomy with relationships and the term-IPC mapping are stored for any time of the main processing.

3.2 Main Processing

The main processing block has two steps of operation. As a primary step, our system use term-IPC mapping repository data for predicting probable IPC related to a given abstract, whereas the next step uses the taxonomy and relationships among IPCs to narrow down the primary selection of IPC. Let us assume that a classifiable abstract contains features, $a = \{fa_1, fa_2...fa_n\}$. Then following section describes both of the steps elaborately.

3.2.1 Predicting Primary Probable IPC

We use repository data for term-IPC mapping for predicting primary probable IPC for a given abstract. We have classifiers that can evaluate the similarity between the classifiable abstract and categories by using weight value of term-IPC mapping. The relevance of an abstract, a to a category, c is defined as

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

where $TF(f,a)$ is the frequency of feature f in the abstract a. If the relevance, Φ_c of an abstract a to a category c is greater than the threshold, θ, then the category is considered as a probable relevant IPC to the abstract. Hence, a set of probable relevant IPC are extracted after applying the first step of main processing block.

The IPCs are organized in sections, classes, subclasses, main groups and subgroups. We have layered model to identify section, class, subclass and IPC as a whole. We found our experiments that sequential identification of section, class, and subclass has positive impact over the results. Our method of predicting probable IPCs can retrieve correct sections, classes and subclasses at the most of time. However, it has limitations retrieving more specific (deeper in the hierarchy) IPCs. Therefore the IPCs by the methods are not considered as final output, rather it is considered as primary probable IPCs. Fig. 2 depicts the overall flow of the methodologies.

3.2.2 Predicting Specific IPC

We use taxonomy of IPCs and their relationships to predict more specific IPCs. In this phase of operation, our system starts off each primary probable IPC and looks for the similarity among the neighbors and similarity in the referenced IPCs of the primary probable IPCs. The similarity is measured starting off primary probable IPC to their neighbors along their paths.

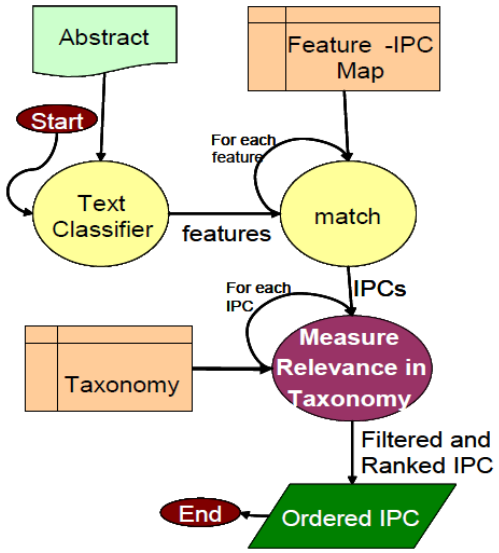

Figure 2: The overall block diagram of our patent mining system which produces ranked list of proposed IPCs for a scientific abstract.

For an example, if primary probable IPC is p and the neighbor is $pc_1, pc_2...pc_n$, then similarity for all paths starting off p to its neighbors is computed. If the improvement in similarity value for a specific path stops, then the deepest neighbor, where the last improvement in similarity occurs is considered as the selected specific IPC of that path.

Ordered list is created from the increased number of IPCs. The most relevant IPCs from the ordered list is further filtered to produce the final output of related IPCs.

Therefore, ontology graph structure of taxonomy and simple relatiohips of reference plays an important role in this step.

4 Result and discussion

There were several systems, which were applied for patent mining task. We got the retrieved result from the NTCIR-7 after formal run, we found that it was very poor [20]. We analyzed our results and sytem and found that this was because we used the taxonomy of IPC of version 8, whereas the formal run of NTCIR-7 used IPC of version 6. There were differences between IPC of version 6 and that of version 8. The millions of patent documents were also based on the IPC of version 6. Moreover, there was no taxonomy of IPC of version 6 available during our experiments through the official website of WIPO.

Again, there is enough room to improve our system focusing on the enhancement of text analyzer part. We are still continuing to improve our system.

5 Conclusion

In this paper, we describe a system to retrieve related ranked IPC for a scientific paper abstract by using ontology of semantic technology. This is a new approach which uses ontology of IPC. Using the semantic technology, our system results relevant IPC quickly. We measured similarities between the sets of features from a scientific paper abstract and a prototype document of a IPC category, which contains all patent documents of that specific category. Although our algorithm is still naive at utilizing the essence of ontology effectively, locality of reference would help the system run faster. Our future target is to enhance our system to be applicable to different versions of IPC using correspondence tables of changes made over versions. We also focus on improving the utilization of ontology and to evaluate the results to measure further precision.

References

[1] S. Adams. Using the International Patent Classification in an Online Environment. *World Patent Information*, 22(4):291–300, 2000.

[2] L. Cai and T. Hofmann. Hierarchical Document Categorization with Support Vector Machines. *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pages 78–87, 2004.

[3] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Using Taxonomy, Discriminants, and Signatures for Navigating in Text Databases. *Proceedings of the International Conference on Very Large Databases*, pages 446–455, 1997.

[4] 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: The International Journal on Very Large Data Bases*, 7(3):163–178, 1998.

[5] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced Hypertext Categorization Using Hyperlinks. *ACM SIGMOD Record*, 27(2):307–318, 1998.

[6] O. Dekel, J. Keshet, and Y. Singer. Large Margin Hierarchical Classification. *Proceedings of the 21st International Conference on Machine learning*, 2004.

[7] S. Dumais and H. Chen. Hierarchical Classification of Web Content. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 256–263, 2000.

[8] C. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. Automated Categorization in the International Patent Classification. *ACM SIGIR Forum*, 37(1):10–25, 2003.

[9] C. Fall, A. Torcsvari, and G. Karetka. Readme Information for WIPO-alpha Autocategorization Training Set, 2002.

[10] C. Fall, A. TOrcsvaxi, P. Fievet, and G. Karetka. Additional Readme Information for WIPO-de Autocategorization Data Set, 2003.

[11] S. Godbole and S. Sarawagi. Discriminative Methods for Multi-labeled Classification. *Lecture Notes in Computer Science*, pages 22–30, 2004.

[12] T. Hofmann, L. Cai, and M. Ciaramita. Learning with Taxonomies: Classifying Documents and Words. In *NIPS Workshop on Syntax, Semantics, and Statistics*, 2003.

[13] N. Kando. What Shall We Evaluate? Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in Patent Searching and Patent Attorneys. In *Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval*, 2000.

[14] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela. Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, 2000.

[15] C. Koster, M. Seutter, and J. Beney. Classifying Patent Applications with Winnow. In *Proceedings Benelearn*, pages 19–26, 2001.

[16] M. Krier and F. Zaccí. Automatic Categorisation Applications at the European Patent Office. *World Patent Information*, 24(3):187–196, 2002.

[17] L. Larkey. Some Issues in the Automatic Classification of US Patents, 1997.

[18] L. Larkey. A Patent Search and Classification System. *Proceedings of the 4th ACM Conference on Digital libraries*, pages 179–187, 1999.

[19] N. Littlestone. Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm. *Machine Learning*, 2(4):285–318, 1988.

[20] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto. Overview of the Patent Mining Task at the NTCIR-7 Workshop. *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 2008.

[21] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Learning Hierarchical Multi-category Text Classification Models. In *Machine Learning-International Workshop Then Conference*, volume 22, page 745, 2005.

[22] M. Ruiz and P. Srinivasan. Hierarchical Text Categorization Using Neural Networks. *Information Retrieval*, 5(1):87–118, 2002.

[23] A. J. C. Trappey, F.-C. Hsu, C. V. Trappey, and L. 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.