Ontology Driven IPC Based Classification of a Research Abstract

Md. Hanif Seddiqui, Masaki Aono *

Abstract—**This paper introduces a method of ontology driven text classification. We retrieve correspondence between research abstract and patent categories (i.e. International Patent Classification or IPC) organized in a taxonomy of an ontology. In the proposed method, we associate each category to a set of related features and their weight-values extracting from the pre-classified huge patent documents. We extract features and the feature-to-feature relations from a research abstract. Then the extracted features of the abstract are mapped with the categories of ontology considering underlying semantics to correspond the proper classifications or categories. We experimented with IPC ontology and huge pre-classified patent documents to classify research abstract and the result shows the strength of our method.**

Keywords: Text Mining, Text Classification, CHI-Square, International Patent Classification (IPC), Ontology, Patent Mining.

1 Introduction

Text classification in the field of information retrieval (IR) is an activity of labeling natural language texts with thematic categories from a predefined set. The standard methods of the machine learning techniques used in text classification usually operate on input documents after they have been transformed into feature vectors $f_1,\ f_2,\ ...,\ f_n\ \in\ D$. Most of the available techniques depend on the syntactic analysis of the features or keywords. They seldom analyze the semantics beneath the text. However, we use the the taxonomy of categories and the undelying semantics in the text for document classification.

We have a large number of IPCs organized by World Intellectual Property Organization (WIPO) and huge number of preclassified patent documents. WIPO maintains IPC within an ontology in XML format [1] having taxonomy of categories and relations such as cross references. The IPC taxonomy consists of about 80,000 categories that cover the whole range of industrial technologies.

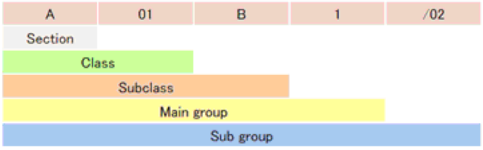

Figure 1: A is a section for ‘Human Necessities’, A01 is class representing ‘Agriculture; Forestry; Hunting; Fishing; etc.’, A01B is subclass which consists of ‘Soil working in agriculture or forestry etc.’, A01B 1/00 is a main group representing ‘Hand Tools’, while A01B 1/02 is a subgroup for ‘Spades;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

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 lowest level (See Fig. 1). The subgroups are even classified into different sublevels. 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.

Classification of a research abstract considering patent categories, like IPC, is a challenging tasks in the field of IR. Many vague and general terminologies are often used to avoid narrowing the scope of the invention [1] in patent documents. Patent documents contain even acronyms and many new technical terminologies [2], which make patent based classification task challenging. Moreover, an abstract usually contains general terminologies to make it understandable by even general people. As a result, indexing of terminologies is not alone sufficient in this field of classification due the above problems. Machine learning based classifier often suggests a number of categories instead of a single one due to the generality of the features. It is a worthy chanllenge of a classification scheme to identify a single relevant category for an abstract within the universe of knowledge. When a document discloses multiple categories of IPCs, the rules of precedence have to be applied in order to determine the final classification with sufficient depth [3]. Therefore, some effective alternative techniques of automatic patent classification is necessary to deal with the challenges.

To overcome the challenges of automatic patent classifi-

*Knowledge Data Engineering Lab, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580, Japan. Email: hanif@kde.ics.tut.ac.jp, aono@ics.tut.ac.jp

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

cation, we introduces a new approach. Our system uses ontology of IPC available in the WIPO official website, and creates model of taxonomy for IPCs. It also develops mapping between features available in patent documents to the IPC of the taxonomy. To classify a research abstract, our system uses the mapping of the feature to IPC to retrieve probable IPCs related to the given abstract. We consider each of the probable IPCs as an anchor point to start off finding more specific categories related to the abstract. Our system further observes the availability of the features of the abstract to the description of neighbors of the probable IPC as decsribed by Seddiqui et. al. [4] in their ontology alignment technique. It refines the probable IPCs taking advantages of the locality of references. Eventually, our system produces more relevant IPC in sufficient depth for a research abstract with the help of ontology and utilizing the technique of ontology alignment. It is capable of generating significantly better results of categorization within short span of elapsed time.

We organize the rest of the paper as follows: Section 2 contains the related works, while Section 3 focuses the text preprocessing of our system. Section 4 describes the main processing unit of our method. We describe our experiments and evaluation in Section 5. Section 6 includes the conclusion and the future directions.

2 Related Works

The document categorization based on the patent classification scheme can be performed in two ways: an algorithm can either flatten the taxonomy of IPC and consider it as independent categories, or can incorporate the hierarchy in the process of categorization. Early patent categorizers chose the former solution, but these were outperformed by the real hierarchical classifiers.

The first hierarchical classifier was developed by Chakrabarti et. al. [5, 6] 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 the feature terms from the noise terms efficiently. They tested the approach on a small-scale subtree of a patent classification consisting of only 12 subclasses organized in three levels [7].

Larkey [8, 9] has created a tool for attributing US patent codes based on a k-Nearest Neighbor (k-NN) approach. The inclusion of phrases (features of multi-word terms) during indexing is reported to have increased the systems precision for patent searching but not for categorization [8], though the overall system precision is not specified.

Kohonen et. al. [10] developed a self-organizing map based PC system. Their baseline solution achieved a precision of 60.6% when classifying patents into 21 categories.

A comprehensive set of patent categorization tests is reported in [11]. 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 [12].

The above listed approaches are difficult to compare due to the lack of a common benchmark patent application collection and a standard patent taxonomy. This lack has been at least partly alleviated with the disclosure of the WIPO and NTCIR [2] document collections. First, the WIPO-alpha English collection was published in 2002 [13], and shortly after the WIPO-de German patent application corpus became publicly available [14]. The creators of the WIPO-alpha collection [1] 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. [15] experimented on the D section (Textile) with 160 leaf level categories and obtained 71.9% accuracy. Rousu et. al. [16] 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 [17] tested their hierarchical SVM-like categorization engine on each section of WIPO-alpha, and obtained 32.4-42.9% accuracy at the maingroup level.

A patent application oriented knowledge management system has been developed by Trappey et. al. [18], which incorporates patent organization, classification and search methodology based on back-propagation neural network (BPNN) technology. Other hierarchical categorization algorithms such as in [19, 20, 21] have not been evaluated on patent categorization benchmarks.

3 Text Preprocessing of Our System

Our system of patent classification includes two major steps for the whole process: preprocessing and the main processing. The preprocessing step contains two independent operations, one is the development persistent memory model of IPC taxonomy from IPC ontology available in XML format [3], and the other is the feature selection.

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

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

The feature selection utilizes the machine learning techniques to normalize a text, to reduce availability of rare words, and to map features to IPCs based on CHI-square technique.

3.1 Creating Taxonomy of IPC

We create our own memory model for the taxonomy along with some simple relations from the IPC ontology available in XML format. We use DOM XML parser to parse the IPC contents. The XML file for IPC contains entryReference tag for referencing other IPCs which are relatively similar, but available in different groups. We parse the entryReference tag as a relationship between IPCs. The relationship in the taxonomy of IPC deals indirect categorization of patent classification. We represent the persistent memory model of IPC taxonomy by directed acyclic graph structure.

3.2 Feature Selection

There are almost one million preclassified English patent documents in a dataset from the year 1993 through 2000. Our text analyzer represents a document as a set of features, $d = \{f_1, f_2, f_3, ..., f_m\}$, where m denotes the number of features that occur in the documents. Moreover, every patent document is associated with a primary IPC. Feature, typically, represents a single or multi-word term having unique meaning together.

3.2.1 Text Normalization

The primary text preprocessing unit parses all the documents and attach POS-tag using stanford POS-tagger [22, 23], removes the standard stop words of English, and stems individual words with porter stemmer. We extract features as both individual words and the multi-words. The normalized features or terms are associated with their corresponding IPC and IPC based term frequency as each documents comes along with primary IPC.

After the text normalization and extraction of feature terms, we use Document Frequency (DF), which is frequently used in Information Retrieval (IR) field, for removing rare terms to reduce feature space and to increase accuracy, and χ^2 based analysis for the processing to assign features to specific categories.

3.2.2 Document Frequency (DF)

The document frequency is represented by the number of documents in which a term occurs. We compute the document frequency for each unique term in the training corpus with the preclassified patent document. Afterwards, we remove the terms from the feature space, whose document frequency is less than the predetermined threshold. We consider the threshold as two. The basic assumption is that rare terms are either non-informative for category prediction or not influential in global performance. In either case removal of rare terms reduces the dimensionality of the feature space. However, we do not use the DF for aggressive term removal because of a widely received assumption in information retrieval, as low (however, not less than threshold) document frequency terms are assumed to be relatively informative.

3.2.3 CHI-Square, χ^2 based Feature Selection

After removing the rare terms, we apply the χ^2 statistic to measure the relationship between term, t and category, c and we compare to the χ^2 distribution with one degree of freedom to judge extremeness. If term t is associated more than one categories, χ^2 is calculated for all of the categories. Using the two way contingency table of a term t, and a category c, where A is the number of times t and c co-occur, B is the number of time the t occurs without c, C is the number of times c occurs without t, D is the number of times neither c nor t occurs and N is the total number of classes. The values $A - D$ are called the observed frequencies (O), and may be arranged in a 2 x 2 contingency table and we calculate the expected frequencies (E) for each table cell according to M. Oakes et al. They defined the Chi-Square using the contingency table as follows:

$$\chi^2(t,c) = \sum_{i,j} \frac{O_{i,j} - E_{i,j}^2}{E_{i,j}} \qquad (1)$$

If χ^2 is greater than 3.84, we can be 95% confident that the word does occur more frequently in one of the two text types. If the ratio A/B is greater than the ratio (A+C)/(A+D), then the word is more typical of the specific category c (a “positive indicator”), otherwise it is more typical of the rest of the category (a “negative indicator”). If the word is classified to be a part of a category c confidently, then the χ^2 value of the word for other category is set to zero. And if χ^2 is not greater than 3.84, we keep the χ^2 value for each of the category. For the calculation to be reliable we must discard any words where E is less than 5). [24]

3.2.4 Taxonomy of the Bag of Words (BOW)

We have taxonomy of IPC derived from the IPC ontology, where IPC are arranged in multiple layers, i.e. section, class, subclass, main group and subgroup categories (See

Fig. 1) and CHI-Square based feature selection is capable of retrieving terms related to a specific category. Therefore, we apply χ^2 based method for selecting features for section of IPC first. In this way, we can get a bag of words or features associated to a particular section. Then, these features of a particular section are futher distributed among its class and so on. After a few iteration, we get a taxonomy of the Bag of Words, which is not necessarily related in terms of meaning, however, they are in the taxonomy for their availability in the related fields.

Although the total preprocessing cost much in computation, however, it is only measured once and kept as a repository. It is reused until the IPC taxonomy or the set of preclassified patent documents are changed. After the preprocessing IPC taxonomy with relationships and the feature-IPC mapping are stored for any time use in the main processing.

4 Main Processing Unit of Our System

The main processing block has three steps of operation. As a primary step, our system process research abstract to be classified. On the next step, we use feature-IPC mapping data for predicting probable IPC related to the given abstract, while on the last step, we use taxonomy alignment to narrow down the primary selection of IPC.

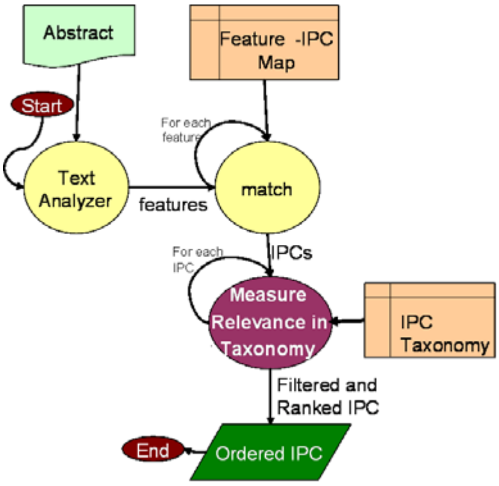

Figure 2: The overall block diagram of our patent mining system which produces ranked list of proposed IPCs for a scientific abstract.

Let us assume that an abstract contains features, $a = \{fa_1, fa_2, ..., fa_n\}$. Then, the following subsections describe the steps quite elaborately.

4.1 Processing Abstract Text

We normalize the abstract to be classified using the techniques described in Section 2.2 and extracts the features. In addition to the normalization process, we discover the lexico-syntactic pattern [25, 26, 27] to detect hyponymy relations of terms. We also use hyponym/hypernym (is-a/is-a-type-of) relation [28] of WordNet for finding relation among features available in the abstract. We develop simple taxonomy of terms, rather than complex ontology learning procedures, to find simple hierarchical relationships only. However, there are some abstract which is too short to extract any relationship among terms. Therefore, at the stage of processing abstract text, we find banch of features, and we may also find a hierarchy of features showing their relationship.

4.2 Predicting Probable IPC

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

$$\Phi_c = \sum_{f \in a} \chi^2(f, c)$$

where $\chi^2(f, c)$ is the weight 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 IPCs to the abstract. Hence, a set of probable relevant IPC are extracted.

4.3 Predicting IPC based on Taxonomies

The previous step considers only the syntactic analysis to predict probable IPCs. However, it has limitations of retrieving accurate IPC as it does not consider the semantic relations. Therefore the IPCs by the steps mentioned above are not considered as a final output, rather it is considered as primary probable IPCs. Fig. 2 depicts the overall flow of the methodologies.

At this point, we have taxonomy of IPC and the taxonomy of the Bag of Words (BOW) associated to IPCs and we already get the probable IPCs. We have a bit semantic (hyponym/hypernym) relations among features of an abstract. To obtain more specific and accurate IPCs, we consider a probable IPC as an anchor point of further finding the more specific IPC. Starting from an anchor IPC in the taxonomy, our system traverses towards the ancestors, siblings IPCs, the descendants and the referenced IPCs for finding the maximum availability of features of the abstract. The possibility of being categorized to an IPC is higher, if the more related features are found. Among the cloud of IPC, the most specific one is the output.

5 Experiment and Evaluation

We experimented considering the whole set of IPC categories, which included about 80,000 categories at section, class, subclass, main-group, and sub-group levels. We considered more than one million preclassified patent documents. We used these preclassified patent documents in the preprocessing stage of our system. Our system was then applied against almost one thousand non-classified research abstract to classify them . We obtained the total dataset used for the workshop of NTCIR-7 held in National Institute of Informatics during mid of December, 2008 in Tokyo of Japan.

The recall and precision of our system is depicted by the recall-precision graph below (See Fig. 3):

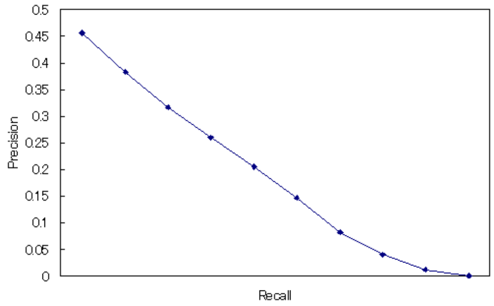

Figure 3: The recall-precision graph of our system, which used a moderate number of test data.

6 Conclusions and Future Work

We introduces rare word reduction, χ^2 based classification and weight generation, association of the bag of words to the taxonomy of categories, and the techniques of extracting relation among features in our system. This approach uses ontology of IPC. Using the semantic technology, our system retrieves relevant IPC quickly and efficiently. Although our algorithm is still naive at utilizing the essence of ontology effectively, locality of reference helps the system run faster. We consider the whole set of categories of IPC. The utilization of ontology of the semantic technology is our novel approach in this field of classification.

Our future target is to improve the phase of ontology learning and to analyze and compare the results comprehensively with those of other systems.

Acknowledgements

This study was supported by Global COE Program gFrontiers of Intelligent Sensingh from Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Hori Information Science Promotion Foundation.

References

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

[2] Kando, N.: 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)

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

[4] Seddiqui, M.H., Aono, M.: Alignment Results of Anchor-Flood Algorithm for OAEI-2008. Proceedings of Ontology Matching Workshop of the 7th International Semantic Web Conference, Karlsruhe, Germany (2008) pp.120–127

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

[6] Chakrabarti, S., Dom, B., Agrawal, R., Raghavan, P.: 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) (1998) pp.163–178

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

[8] Larkey, L.: Some Issues in the Automatic Classification of US Patents (1997)

[9] Larkey, L.: A Patent Search and Classification System. Proceedings of the 4th ACM Conference on Digital libraries (1999) pp.179–187

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

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

[12] Koster, C., Seutter, M., Beney, J.: Classifying Patent Applications with Winnow. In: Proceedings Benelearn. (2001) pp.19–26

[13] Fall, C., Torcsvari, A., Karetka, G.: Readme Information for WIPO-alpha Autocategorization Training Set (2002)

[14] Fall, C., TOrcsvaxi, A., Fievet, P., Karetka, G.: Additional Readme Information for WIPO-de Autocategorization Data Set (2003)

[15] Hofmann, T., Cai, L., Ciaramita, M.: Learning with Taxonomies: Classifying Documents and Words. In: Proceedings of NIPS Workshop on Syntax, Semantics, and Statistics. (2003)

[16] Rousu, J., Saunders, C., Szedmak, S., Shawe-Taylor, J.: Learning Hierarchical Multi-category Text Classification Models. In: Proceedings of the 22nd International Conference on Machine Learning. Volume 22. (2005) pp.744–751

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

[18] Trappey, A.J.C., Hsu, F.C., Trappey, C.V., C.-I., L.: Development of a Patent Document Classification and Search Platform Using a Back-propagation Network. Expert Systems with Applications **31**(4) (2006) pp.755–765

[19] Dekel, O., Keshet, J., Singer, Y.: Large Margin Hierarchical Classification. Proceedings of the 21st International Conference on Machine learning (2004)

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

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

[22] Toutanova, K., Manning, C.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000) (2000) pp.63–70

[23] Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (2003) pp.173–180

[24] Oakes, M., Gaaizauskas, R., Fowkes, H., Jonsson, A., Wan, V., Beaulieu, M.: A method based on the chi-square test for document classification. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2001) pp.440–441

[25] Hearst, M.: Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics-Volume 2 (1992) pp.539–545

[26] Moldovan, D., Girju, R., Rus, V.: Domain-Specific Knowledge Acquisition from Text. Proceedings of the 6th Applied Natural Language Processing (ANLP-2000) Conference (2000) pp.268–275

[27] Buitelaar, P., Olejnik, D., Sintek, M.: A Protege Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. Lecture Notes in Computer Science (2004) pp.31–44

[28] Hearst, M.: Automated Discovery of WordNet Relations. WordNet: An Electronic Lexical Database (1998) pp.131–151