

Use of Ontology in Text Classification

Md. Hanif Seddiqui and Masaki Aono

Toyohashi University of Technology, Aichi, Japan
hanif@kde.ics.tut.ac.jp, aono@ics.tut.ac.jp

Abstract. This paper introduces a method of ontology based text classification. The concept of ontology is associated with the related words and their weights from the pre-classified documents as a learning stage. In the main process, the words and their mutual relations are extracted from the target documents. Then the target document is mapped with concepts of ontology. We experiment with International Patent Classification (IPC) ontology and huge pre-classified patent documents to classify research abstract.

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

1 Introduction

Various keyword-based technologies are suggested for document categorization nowadays. However, we use the taxonomy of categories for document classification. In this regards, we use International Patent Classification to categorize research abstract.

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 format¹ 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.

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 [1]. Patent document contains even acronyms and many new technical terminologies [2], which make patent mining task challenging. Moreover, indexing of terminologies is not sufficient in patent mining system as

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



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 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 [3]. 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 research abstract with the help of ontology and utilizing the techniques of ontology alignment. Theoretically, it is capable of generating significantly better categorization results within short elapsed time.

We organize the rest of the paper as follows: Section 2 focuses the text preprocessing, while Section 3 describes our main processing unit. We discuss our system in Section 4.

2 Text Preprocessing

Our system includes two major steps for the whole process: preprocessing and the main processing. The preprocessing steps contains two independent operations, one is the development of IPC taxonomy from IPC ontology available in XML format ², and the other is the machine learning techniques to normalize the text and term to category relationship, i.e. term to IPC relationship in our system.

² http://www.wipo.int/classifications/ipc/en/download_area/20080101/xml/ipcr_scheme.20080101.zip

2.1 Creating Taxonomy of IPC

We create the taxonomy along with some simple relations from the IPC ontology data available in XML format. We used DOM XML parser to parse the IPC contents. The XML file for IPC contains entryReference tag for referencing 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 deals many indirect categorization of patent classification. We represent the in-memory model of IPC taxonomy by direct graph structure.

2.2 Feature Selection

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

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

After the text normalization and extraction of terms, we use Category Frequency (CF), which is almost similar to widely used Document Frequency (DF) in Information Retrieval (IR) field, for removing rare terms to reduce feature space and to increase accuracy, and Chi-square based analysis for the primary processing to assign features to specific categories.

Category Frequency (CF) Category frequency is similar to the document frequency. We consider all documents associated with a specific IPC as a collective document, or a unique document of that category. Then, the Category Frequency is the number of category in which a term occurs. We compute the category 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 CF for aggressive term removal because of a widely received assumption in information retrieval, as low (however, not less than threshold) Document Frequency (we are using CF instead) terms are assumed to be relatively informative.

CHI-Square 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 2×2 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:

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

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). [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 Figure 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 further distributed among its class and so on. After a few iteration, we will 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 and the set of preclassified patent documents are changed. After the preprocessing IPC taxonomy with relationships and the term-IPC mapping are stored for any time of the main processing.

3 Main Processing

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 term-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.

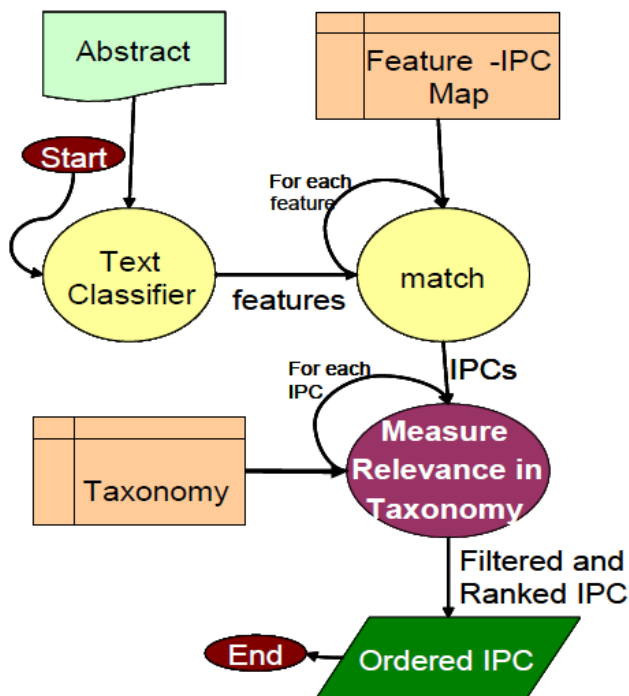


Fig. 2. The overall block diagram of our patent mining system which produces ranked list of proposed IPCs for a scientific abstract.

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

3.1 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 [5–7] to detect hyponymy relations of terms. We also use hyponym/hypernym (is-a/is-a-type-of) relation [8] of WordNet for

finding relation among words 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 bunch of features, and we may also find a hierarchy of features showing their relationship.

3.2 Predicting 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

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

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 IPC to the abstract. Hence, a set of probable relevant IPC are extracted.

3.3 Predicting IPC based on Taxonomies

The previous step only consider 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 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 words of target document. To obtain more specific and accurate IPCs, we consider a probable IPC as an anchor point of further finding the more specific IPC based on the availability of the features of target document. 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 target document. The possibility of being categorized to an IPC, if the more related words are found. Among the cloud of IPC, the most specific one is the output.

4 Discussion

This paper describes the modified techniques of our patent mining system of the NTCIR-7. We enhance our system on several point of view. We introduces rare word reduction, χ^2 based classification and weight generation, taxonomy of the bag of words, and the lexico-syntactic pattern based word relation. This

approach uses ontology of IPC. Using the semantic technology, our system retrieves relevant IPC quickly. Although our algorithm is still naive at utilizing the essence of ontology effectively, locality of reference would help the system run faster.

References

1. Fall, C., Töröcsvári, A., Benzineb, K., Karetka, G.: Automated Categorization in the International Patent Classification. *ACM SIGIR Forum* **37**(1) (2003) 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) 291–300
4. 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) 440–441
5. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th conference on Computational linguistics- Volume 2* (1992) 539–545
6. Moldovan, D., Girju, R., Rus, V.: Domain-Specific Knowledge Acquisition from Text. *Proceedings of the 6th Applied Natural Language Processing (ANLP-2000) Conference* (2000) 268–275
7. 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) 31–44
8. Hearst, M.: Automated Discovery of WordNet Relations. *WordNet: An Electronic Lexical Database* (1998) 131–151