THE STUDY OF PATENT PRIOR ART RETRIEVAL USING CLAIM STRUCTURE AND LINK ANALYSIS

Fu-ren Lin, Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan, R.O.C., frlin@mx.nthu.edu.tw

Feng-mei Huang, Institute of Technology Management, National Tsing Hua University, Hsinchu, Taiwan 300, R.O.C., jalumay@gmail.com

Abstract

Prior art retrieval plays an important role in patent examination. If a patent examiner can quickly and accurately retrieve prior art for an application patent, s/he will be able to efficiently and effectively judge the novelty of an application patent, and in turn, avoid hampering the technology development of the application domain. Moreover, in order to avoid losing tangible and intangible assets caused by patent infringement, companies need to search patent prior art before doing technology development. However, the increasing number of patents is not seen only in quantity but also in complexity. Bewaring the emerging need of prior art retrieval from large patent databases, the main objective of this study is to develop a methodology to effectively identify patent prior art. The developed prior art search framework firstly transforms claim sentences into the claim tree to represent the hierarchy of claimed elements, and then use the specification in patents to expand the terms in nodes of the claim tree. Finally, we calculate the similarity of claim trees and use the threshold to identify the most relevant patents as prior art. The prior art identified by the generated system was compared in recall, precision, and F-measure with those identified by the patent examiner. The results show the effectiveness of the proposed prior art retrieval system.

Keywords: Patent retrieval, Prior art, Claim hierarchy, Link analysis.
1. INTRODUCTION

According to USPTO (United States Patent and Trademark Office, http://www.uspto.gov/) patent statistics report, since 1994, the number of utility patent applications in US has exceeded 100 thousands, and about 50 thousand patents among them were granted every year. The increasing number of patents is not only seen in quantity but also in complexity. This explosive growth of patents indicates the fast growth of inventions, but it also causes the delay of the patent examination process. If a patent office wants to keep promised patent review cycle time, it may face the difficulty in maintaining patent quality. How a patent office balances the cost of handling the increasing number of patent applications and the quality of granting patents is an important issue. Nowadays, some attempts have been used to approximate to this goal, for example, hiring more well-trained examiners, using internet and information retrieve (IR) techniques, etc. Specifically, in 2006, USPTO proposed the 2007-2012 Strategic Plan1 and launched a new cooperation plan with Japan called Pilot Patent Prosecution Highway2. The main objective of these plans is to reduce the workload of approving patents and to improve the patent quality.

According to Title 35, United States Code, Section 102 states that “a person is not entitled to a patent if the invention was known or used by others in this country, or was patented or described in a printed publication in this or a foreign country” before the date of invention by the applicant for the patent. A prior art of an application patent is an invention that was patented or described in a printed publication in the same country or foreign countries before the application patent. Moreover in the United States, applicants have the duty to inform Information Disclosure Statement (IDS) about the prior art of the invention which they are aware of. However, if an applicant does not specify prior art to the IDS, when later on someone else proposes “post-grant opposition” to provide the evidence to prove that the applicant is not honest, the penalty will be triple. In considering the consequence of filing incomplete prior art in applying patents, it is very important to identify prior art effectively and efficiently.

In the recent survey on the challenge of searching prior art, Barrish et al. (2004) proposed three levels of influence while finding the prior art: (1) invalidity, to seek the references that could threaten the invention; (2) blocking patent, seeking the identical patent which claim covers the invention; (3) clearance, searching the products that are being finalized and commercialized. These three will make a patent invalid or threaten commercial products. Therefore, a company cannot secure and make profits from its inventions without monitoring emerging technologies in the world.

Bewareing the emerging need of prior art retrieval from large patent databases, the main objective of this study is to develop a method to identify patent prior art. The claim of the patent is a key factor that determines the patentability after comparing the prior art. In other words, focusing on claim, we can increase the precision of the prior art search. In order to identify the subject matters of the patent, we should construct the claim first. This is the process that was usually ignored by many existing patent tools which were usually using the abstract and specification of a patent to search.

2. LITERATURE REVIEW

Nowadays there are more and more workshops talking about patents, for example, NTCIR (Kando, 2000) discussed the Patent IR Challenge (PIC) in patent searching and patent attorneys, challenges like high recall, high precision and document structure, etc. SIGIR workshop also provided the session to discuss the patent retrieval on searching, cross language and ontology, etc. Due to the similarity of goals of this study with the information retrieval discipline; that is, recall and precision, we discuss patent search papers in these two points of view in Subsection 2.1. Then, Subsection 2.2 reviews the techniques for claim tree construction.

1 http://www.uspto.gov/web/offices/com/strat2007/
2 http://www.uspto.gov/web/offices/pac/mpep/documents/0900_904_02.htm#sect904.02
2.1 Patent Search

Using keywords to search patents is just like searching general documents in which it must handle the query expansion problem. However, this problem is more critical in patent search than in general document search, especially in patent prior art search. Larkey (1999) proposed some ideas to expand the query terms to increase the recall of patent search. He suggested using phrase co-occurrence and dictionary to find the relevant words. Inoue et al. (2000) proposed a similar concept that uses probabilistic method and Hierarchical Bayesian Clustering (HBC) to calculate the similarity and relevance of words between patents. Furthermore, Lim et al. (2004) used ontology to expand query terms by adopting international-standard classification UNSPSC to construe the ontology of patent information for the timepiece industry. These methods mentioned above deal with query expansion to gain higher recall rate; however, we still need other ways to help increase the precision of patent retrieval. Mase et al. (2005) proposed a two-stage patent retrieval method, in which the first stage uses the similar approach to expand the query, and the second stage focuses on claim parsing to re-weigh the query words and then identify the top 1000 patents. Mase’s work is much like this study, but they ignored the importance of citation and claim structure. Citations in some sense provide much information inside the patents; e.g., citations present the importance of patents, and moreover, citations may denote the similarity between inventions. Based on the similarity between patents, we can save a lot of computing time to compare the similarity. Claim structure can provide more structure similarity match than solely using keywords.

2.2 Claim Construction

East (1995) proposed the best way to track the claim in *IEEE AES system magazine*. Based on his suggestion, we need to divide independent claims into different parts according to features in the claim, and then we can compare these features in the whole claim in a chart, called “claim chart”. This method is widely accepted by many people who want to construe the claim.

Sheremetyeva et al. (1996) used templates and POS-tags to parse patent claims. They identified some verbs as predicates. Based on these verbs in the claim, they segment and label words according to their elements (features), and then form a *case-role*. Moreover, the comparison of the similarity and co-occurrence of features between these and other *case-roles* to construct a conceptual schema tree. Furthermore, Sheremetyeva (2003) used *supertags* to improve the POS tagging, and showed the significant performance gains.

Shinmori et al. (2003) proposed the idea of using Rhetorical Structure Theory (RST) to parse a patent claim. According to Marcu (1997), RST relies on “cue phrases” in implementing algorithms to discover the valid RST trees for a single document. This work is reasonable based on the assumption that an author has a fixed writing style. In spite of the different inventors of various patents, RST still makes sense on analyzing patent claim because claims follow a common agreed grammar style. Shinmori et al. (2003) defined many cue phrases that can be used for segmenting long claims and establishing relations among segments or spans. After identifying the elements of a claim, they align the claim with the description of the patent. The alignment step made the claim more readable. In this step, they denoted \( s(b_1, b_2) \) to represent the degree of similarity which \( b_1 \) is a core element (claim sentence) and \( b_2 \) is an extracted sentence in patent description. \( s(b_1, b_2) \) is set to 3 if \( b_1 \) and \( b_2 \) are completely equal except comma if any. \( s(b_1, b_2) \) is set to 2 if the head word for \( b_1 \) and \( b_2 \) are declinable and the basic forms are equal. \( s(b_1, b_2) \) is set to 1 if \( b_1 \) and \( b_2 \) contain a common word.

In this study, we follow Sheremetyeva’s method (1996) to implement the claim tree construction because it’s easier and time saving to use existing POS tagging and claim grammar rather than implementing many RST trees to build the claim structure. Distinct from Sheremetyeva’s method, this study constructs a claim tree based on the description of independent and dependent claims. Thus, in the claim parsing stage, we don’t need to take too much effort to compute the similarity or co-occurrence of the elements (features) in the claim; instead, all we need is a little grammar cue. For example, if it is the same element (feature) mentioned previously, the sentence for the element (feature) will put *the* or *the said*; otherwise, it will put indefinite articles, such as *a* or *an*. After parsing claim sentences, we agree with Shinmori to use alignment to make claim more readable.
However, we want to choose relevant words in the description since the technique of dividing the similarity that Shinmori adopted is too rough to achieve the goal.

3. PATENT PRIOR ART RETRIEVAL SYSTEM

This study aims to design a patent prior art retrieval system to retrieve the most relevant patents called prior art from a patent database, e.g., USPTO, corresponding to a specified patent as illustrated in Figure 1. It consists of four tasks: (1) candidate prior art collection, (2) claim tree construction, (3) claim tree link analysis, and (4) prior art identification. In order to maintain the patent quality, patent offices need to find patent prior art as broad as possible. If a company wants to commercialize its products, it needs to search patent prior art widely as well. This topic has been widely discussed in many papers, so that we can adopt existing tools to find the potential prior art. However, the volume of patents is too much for human to read. Thus, we transform the claim statement as a tree hierarchical structure called claim tree in order to facilitate the automated search for patent prior art. Through claim hierarchies, we can identify the structure of the invention to determine the patentability; in other words, focusing on the claim tree can help us increase the precision of retrieval.

![Figure 1. The system framework for patent prior art retrieval](image)

### 3.1 Candidate Prior Art Collection

The collection of pattern prior art is the subset of the patent database to relieve the computational effort on transforming all patent claims into claim trees. Many search tools can be used to find the prior art of a patent. Google, for example, is the most popular searching engine that stores billions of Web pages. Google provides the advanced patent search in U.S. patent database. Alta Vita presents more patent related documents, and Vivisimo and Clusty retrieve clusters of different domains of patents according to keywords. Patent analysis tools such as Delphion, Derwent, Dialog and Nerac Inc., present analytical results with visualization. These techniques can help us first widely search the candidate prior art and maintain the recall rate.

### 3.2 Claim tree construction

The transformation of claim statements to a claim tree structure consists of three tasks: (1) parsing claim sentences, (2) extracting the relations between elements to build a claim tree, and (3) instantiating nodes on claim trees with relevant terms in specification.

This study adopts claim grammar templates (USPTO patent rules) and regular expression to construct the claim structure. Some claim grammar rules are elaborated as follows. First, claim sentence should start with a capital letter and end at a period, and it could be divided into three parts: preamble,
transitional phrase and body. Second, when introducing the element, it must accompany the indefinite article (“a” or “an”), and the subsequent mention of the element is modified by the definite article or by “said” or “the (said)”, such that the later one is the same as the former one. Based on these rules, a claim has the basic writing types, such as combination, Jepson, Markush and means-plus-function types. The combination type is the basic one for a claim, which looks like “A chair [preamble] comprising [transitional phrase]: a lag, and a back [body].” The Jepson type usually uses a template, such as “A chair comprising a leg, and a back [preamble], the improvement comprising [transitional phrase] a wheel on the leg [body].” In the Jepson type, the transitional phrase can be “the improvement comprising” or “characterized in that (by)”. The preamble part of the Jepson format denotes the known prior art, and the invention feature is clearly limited in body. In this manner, Jepson type saves time to argue about the prior art. Markush type is typically used in chemistry or pharmacy, where the body is selected from the group of “functionally equivalent” entities, for example, “A chemical object [preamble] selected from the group consisting of [transitional phrase] x, y and z [body]”. Sometimes, an applicant wants to describe an ”effect” rather than the components of apparatus, so that s/he can use means-plus-function. When expressing means-plus-function, the claim sentence has such words as “means for” or “means…for….” to describe the effect.

Besides, using regular expression, we also use the NLP tool developed by Stanford Nature Language Processing Group to tag words and find the noun phrases. The Stanford Tagger uses maximum entropy (Kristina et al. 2000) and cyclic dependency tree (Kristina et al. 2003) to calculate the possibility of tags, and then choose the pertinent tag. The biggest constraint of this study is on training data, and Stanford Tagger uses Penn Treebank (Beatrice, 1990) as the tagging dictionary.

After parsing the claim, we can identify four relations between elements in a claim tree. They are comprising, linking, preposition, and verb relations. Comprising relation denotes ingredients of the elements, Linking relation denotes the equivalence of the elements, Preposition relation contains the term of “of”, “with”, or other prepositions between the elements, and Verb relation denotes a verb between two elements. However we don’t consider the relation under the adjective clause because adjective clause contains too much detailed information that would make the hierarchy complicated.

After building the claim tree, the nodes on claim trees can be instantiated by relevant terms in specification. Because inventors usually use broad terms to describe the subject in order to enlarge the target and prior art hierarchies, where a node represented by a vector is denoted by a number. Each node contains the vector of terms which are extracted from claim sentence and specification. First, nodes in the target hierarchy are labeled as $n_i$, as shown in Figure 3. Second, cosine similarity shown in Eq(2) is adopted to calculate the similarity between nodes in the prior art hierarchy labeled as $m_i$.

$$I_{xy} = p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

### 3.3 Claim Tree Link Analysis

In the link analysis of two claim trees, we calculate the similarity of nodes and structures of claim trees between a target pattern and a candidate patent prior art. We propose a novel tree comparison method to compare the similarity between two tree hierarchies. Figure 3 depicts the links between the target and prior art hierarchies, where a node represented by a vector is denoted by a number. Each node contains the vector of terms which are extracted from claim sentence and specification. First, nodes in the target hierarchy are labeled as $n$, as shown in Figure 3. Second, cosine similarity shown in Eq(2) is adopted to calculate the similarity between nodes in the prior art hierarchy labeled as $m_i$. 
and those in the target hierarchy. In Eq(2), \( n_i \) and \( m_j \) denote nodes in target and prior art hierarchy, respectively. \( W_{ni} \) and \( W_{mj} \) depict the weights of the \( j \)th term of the vectors in \( n_i \) and \( m_j \). Then, if the similarity of two vectors is above the threshold, the corresponding node will be linked.

\[
\cos_{n_i, m_j} = \frac{\sum_{j} w_{ni,j} w_{mj,j}}{\sqrt{\sum_{j} w_{ni,j}^2} \sqrt{\sum_{j} w_{mj,j}^2}}
\]  

(2)

Take Figure 2 as an example, node 1 in the prior art hierarchy is labeled to \( n_2 \) which is called \( m_2 \) and node 2 in the prior art hierarchy is labeled to \( n_4 \) which is called \( m_4 \). Note that a node in the prior art hierarchy can be labeled by more than one links, and vice versa. Third, we build the adjacency matrices of hierarchy for a target patent and a prior art. In the matrices, we use the binary value zero and non-zero to denote the relation between each node in the hierarchy (Figure 3). The order of the column and row of a prior art hierarchy follows the sequence of target hierarchy. In Figure 3(b), function \( l \) will transform nodes in the prior art hierarchy to nodes of the target hierarchy. The values of matrix denote the closeness between two nodes; for example, in Figure 3(a), the connection from \( n_1 \) to \( n_2 \) is much closer than that from \( n_1 \) to \( n_3 \), and in this way, the weight from \( n_1 \) to \( n_2 \) is higher than that from \( n_1 \) to \( n_3 \). This connection weight \( C \) will be considered in Eq(3).

Finally, \( \text{AND} \) operator is used to calculate the structure similarity between a target hierarchy and a prior art hierarchy. Take Figure 3 as an example, we can find the similar structure from \( n_2 \) to \{\( n_4, n_5 \)\} and that from \( n_4 \) to \( n_5 \) between a target hierarchy and a prior art hierarchy.

**Figure 2.** The link between the target hierarchy and prior art hierarchy

**Figure 3.** Link adjacency matrixes

### 3.4 Prior Art Identification

After finishing link analysis, we sum the relevance score between a target patent and a prior art patent. Besides, counting the cosine similarity, we also consider the factor of connection weight \( C \), which denotes the closeness of connection as \( C_{\text{prior art}} \) and \( C_{\text{target}} \). In Eq(3), where \( i \) denotes the number of the
similar hierarchies between the prior art hierarchy and the target hierarchy. For example, in Figure 4, the similar structure of target claim hierarchy are \( \{n_2, n_4\}, \{n_2, n_5\} \) and \( \{n_4, n_5\} \) and that of the prior art claim hierarchy are \( \{m_2, m_4\}, \{m_2, m_5\} \) and \( \{m_4, m_5\} \), thus, the \( R \) is sum of the

\[
\sum_i C_i \left( \sum \cos(n_i, m_i) \right)
\]

(3)

4. THE EVALUATION OF PRIOR ART RETRIEVAL SYSTEM

This study integrates the proposed techniques and develops a system to retrieve a target patent’s prior art. The output of the system is a list of prior art retrieved from the candidate prior art collection. In evaluating the performance of the system, we treat the prior art which is listed by an examiner as the correct answer. We then compute the precision, recall and F-measure by comparing the prior art retrieved by the system with the prior art identified by the examiner.

4.1 Data Collection

To evaluate the system, we first chose the patents which are related to nanoimprint. Nanoimprint is an important technology to develop small but efficient wafers. We selected the patent No.7137803, “Fluid pressure imprint lithography”, which is filed in 2002 as the target patent. The example claim tree for patent No.7137803 is illustrated in Figure 4.

We check the PAIR (Patent Application Information Retrieval) developed by USPTO to obtain the examination record of the target patent, including the examiner’s search strategy. Examiner’s search strategy lists the process of search strategy to find the prior art according to a target patent. These strategies contain keywords related to the invention or classification having similar inventions. This strategy information is very important for this system to find candidate prior art since a patent examiner is an expert to a specific domain and his or her suggestions are reliable as well. Thus, we extracted candidate prior art using keywords, such as mold, nanoimprint, surface with metal, and retrieved 4,388 patents as the candidate prior art collection in total, excluding the patents which have incomplete content in the html patent version.

4.2 Experimental Design

Because it’s time consuming to transform all patents into claim trees, in the experiments, we randomly selected 144 patents out of 4,388 patents as the candidate prior art, including 28 prior art listed by the examiner. Then, we transformed these patents into claim trees, and invited a domain expert to examine these claim hierarchies. The construction of claim hierarchies is improved based on a domain expert’s input.

The parameters used for experiments are set as follows. In instantiating claim terms with specification terms, we set threshold \( \mu \) for \( TF \) and \( MI \) to select the important words in the specification and determine which terms should be aligned in the claim hierarchy, respectively. The weight of a term in the main element of claim hierarchy is set to 3 to emphasize its high impact on specifying patent claim. The weight of a term with \textit{linking relation} with the main element is set to 2 due to its relation with main elements. The weight of a word in the specification is set to 1 due to its relation with the main element. In link analysis and prior art identification tasks, we give an equal weight of starting score of comprising, verb and preposition relations, but the weight of their child nodes decreases according to the type of relation; for example, the \textit{verb relation} decreases 0.1 and \textit{preposition relation} decreases 0.3. We use different thresholds of \textit{cosine similarity} to evaluate the recall, precision and F-measure of the proposed system.
4.3 Experimental Results and Discussion

This study uses different thresholds of cosine similarity to retrieve the prior art in evaluating results. The formulas for recall, precision, and F-measure are shown in Eq(4), Eq(5), and Eq(6), respectively, where the relevant documents denote the prior art identified by a patent examiner, and the retrieved documents denote the prior art retrieved by the system. 28 out of 144 candidate prior art were identified by the patent examiner as prior art. The results are listed in Table 1.

\[
\text{Recall} = \frac{\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}}{\{\text{relevant documents}\}} \tag{4}
\]

\[
\text{Precision} = \frac{\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}}{\{\text{retrieved documents}\}} \tag{5}
\]

\[
F - \text{measure} = \frac{2(\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \tag{6}
\]

Table 1. The experimental results in recall, precision and F-measure

<table>
<thead>
<tr>
<th>Performance</th>
<th>(\mu + 4\sigma)</th>
<th>(\mu + 3\sigma)</th>
<th>(\mu + 2\sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.29</td>
<td>0.79</td>
<td>0.93</td>
</tr>
<tr>
<td>Precision</td>
<td>0.62</td>
<td>0.42</td>
<td>0.26</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.39</td>
<td>0.56</td>
<td>0.41</td>
</tr>
</tbody>
</table>

The result shows that the system can retrieve about 93% of prior art with low threshold, such as \(\mu + 2\sigma\). However, the precision is relatively low, such that additional human efforts are needed to identify prior art. Nevertheless, a patent examiner can save time from verifying most of the less relevant patents retrieved by other methods using keywords or citations. These outcomes can be explained as follows. First, this methodology involves natural language processing. NLP nowadays also needs to do a certain degree of modification to suit the claim grammar. Second, the principle for an examiner to decide a prior art considers “novelty” and “non-obviousness”. Novelty is estimated through one to one comparison with the target patent and prior art, and this comparison process is similar with the methods used in this study. However, the assessment of non-obviousness is not as simple as that of novelty. When deciding the obviousness, the examiner takes more than one candidate prior art patents together and compare them with the target patent. In the experiment, the technique of “imprint
lithography” is still very new for patent database filed in 2002. Therefore, the principle of “non-obviousness” is more important than “novelty” for the prior art of patent No.7137803.

6. CONCLUSIONS

In the study, we propose a novel technique to retrieve a target patent’s prior art. In parsing the patent, many studies only extract words from patents. However, this study builds a claim hierarchy for each patent according to the extracted components from the patent’s claim and specification. The weights for terms from different locations of a patent, e.g., claim or specification, have different scales in order to differentiate their importance in similarity analysis between patents. The comparison between a target patent and candidate prior art through claim hierarchies generates a set of prior art with different degrees of similarity.

The proposed system conducts four main tasks: (1) candidate pattern prior art collection, (2) claim tree construction, (3) claim tree link analysis, and (4) prior art identification. The function of candidate pattern prior art collection identifies the candidate prior art using the existing patent search tools, such as Delphion, Google patent, etc. Claim tree construction transforms a patent’s claim statements to a claim hierarchy, and identifies four types of relation between claimed elements: comprising, linking, verb and preposition. Claim tree link analysis expands the node’s constituent terms of a claim hierarchy using link analysis. Prior art identification locates the prior art which is the collection of the most relevant patents with the target patent.

This study based on the transformed claim trees to identify prior art contributes to the practice of patent search in the following dimensions. First, it can reduce patent examiners’ or analysts’ cognitive loads taken in retrieving prior art. An examiner can balance the cost of dealing with the increasing number of patent applications and the quality of granting patents through claim hierarchy comparison. A company can use claim hierarchy to identify prior art and speed up the time to market. Second, if we collect many claim hierarchies on one specific technology domain, we can efficiently monitor the essential claim structure based on these claim hierarchies. Third, we can identify the similar patent structure of different companies to facilitate the patent portfolio management.

References


Kando, N. (2000). What Shall We Evaluate? -- Preliminary Discussion for the NTCIR.


