On the Role of Classification in Patent Invalidity Searches

Christopher Harris[1], Steven Foster[2], Robert Arens[3], Padmini Srinivasan[1,3]

[1]Informatics Program
The University of Iowa
Iowa City, IA 52242
christopher-harris@uiowa.edu

[2]Global News Intelligence
Montpelier, VT 05667

sfoster@epigeny.com

[3]Computer Science Department
The University of Iowa
Iowa City, IA 52242
{robert-arens, padmini-srinivasan} @ uiowa.edu

ABSTRACT

Searches on patents to determine prior art violations are often cumbersome and require extensive manpower to accomplish successfully. When time is constrained, an automatically generated list of candidate patents may decrease search costs and improve search efficiency. We examine whether semantic relations inferred from the pseudo-hierarchy of patent classifications can contribute to the recognition of related patents. We examine a similarity measure for hierarchically-ordered patent classes and subclasses and return a ranked list of candidate patents, using a similarity measure that has demonstrated its effectiveness when applied to WordNet ontologies. We then demonstrate that this ranked list of candidate patents allows us to better constrain the effort needed to examine for prior art violations on a target patent.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval Language—*Retrieval models*

General Terms

Algorithms, Experimentation, Theory.

1. INTRODUCTION

1.1 Background and Motivation

One of the primary responsibilities of a patent examiner is to scrutinize a target patent for prior art violations. Patent searches involve obtaining a list of all candidate patents and concepts that could potentially infringe upon a target patent, and then manually refining the list, which is both laborious and prone to errors of omission. However, the resources available for patent searches are frequently constrained by limitations of time or manpower; hence the need for a ranked list of most likely patent violators.

Patent-issuing bodies such as the European Patent Office (EPO) and the United States Patent and Trade Office (USPTO) manually classify each patent application into one of many class/subclass combinations (called a *classification*) based on the patent's intended use[1]. Subclasses serve as a more granular categorization of a particular class. The USPTO, for example, classifies each patent into at least one of approximately 470 classes and 163,000 subclasses [9].

Patent classes and subclasses categorized by the USPTO are hierarchical, though a patent's classification in the USPTO system appears as if there are only two levels (top-level class and most-distinct subclass). Because classes and subclasses are often nested, the extended hierarchy can have as many as 14 distinct subclass levels for a given class. Figures 1 and 2 illustrate the differences between how the USPTO defines these relationships in the class descriptions and how the class/subclass hierarchy is truly represented in a tree structure. For example, class 521 (synthetic resins) has a number of subclasses, one of which is 50 (cellular products or processes etc.). This subclass itself has a number of subclasses, e.g. 82 (processes of forming a cellular product etc.). Even though 50 and 82 are in a super/subclass relationship, both represented as subclasses of 521, i.e. 521/50 and 521/82. Both are fully-qualified patent classifications used by USPTO.

Complicating this further is the dynamic nature of patent classifications. Earlier investigation by Larkey [4] found that a single subclass can have up to 2000 patents, but the USPTO attempts to limit the patents in a single subclass to no more than 200 by creating additional subclasses. Periodic reviews of the classification system by the USPTO often results in the restructuring of many subclasses by further dividing, merging, or eliminating subclasses. Additionally, new inventions may require an entirely new set of classes and subclasses to be introduced to accurately describe the invention's intended use.

In this paper, we focus exclusively on the USPTO's patent classification system (USPC) and how USPC classes and subclasses can be used to produce a ranked list of candidate patents. We chose to constrain the patents we used in this study to a single concentration (chemistry), since this more accurately reflects measures a patent examiner would undertake. We focused on those classes determined by CAS (Chemical Abstract Service) to be related to chemistry [8], but our methodology could be easily extended to other fields.

PaIR'09, November 6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-809-4/09/11...$10.00

[1] The EPO uses the International Patent Classifications (IPC) established by WIPO (World Intellectual Property Organization). The IPC group/subgroup follows a similar hierarchical structure as the USPC class/subclass.

Each patent is given at least one mandatory classification, called an Original Classification (OR) by the USPTO. This OR is determined by the controlling claim of the patent. Most patents are also given one or more optional classifications, known as a cross-reference classification (XR). In this paper, we refer to ORs as *primary classifications* and XRs as *secondary classifications*. Additionally, each patent may contain one or more *cited references*, which are prior art references cited by the USPTO during a patent examination or the patent's inventor prior to submission. These contain earlier patents and publications disclosing inventions deemed similar to the patent investigated.

CLASS 521: SYNTHETIC RESINS OR NATURAL RUBBERS -- PART OF THE CLASS 520 SERIES

...

50. CELLULAR PRODUCTS OR PROCESSES OF PREPARING A CELLULAR PRODUCT, E.G., FOAMS, PORES, CHANNELS, ETC.: This subclass is indented under Class 520, subclass 1.

...

82. Process of forming a cellular product subsequent to solid polymer formation in the presence of a stated ingredient, noncellular composition capable of forming a cellular product and containing a stated ingredient, or process of preparing same: This subclass is indented under subclass 50.

...

94. Ingredient is a nitrogen containing compound: This subclass is indented under subclass 82.

95. Nitrogen compound contains a nitrogen atom bonded to a nitrogen or oxygen atom: This subclass is indented under subclass 94.

Figure 1. Example of the USPC Class (521) and its selected subclasses as represented in the USPTO Classification Manual

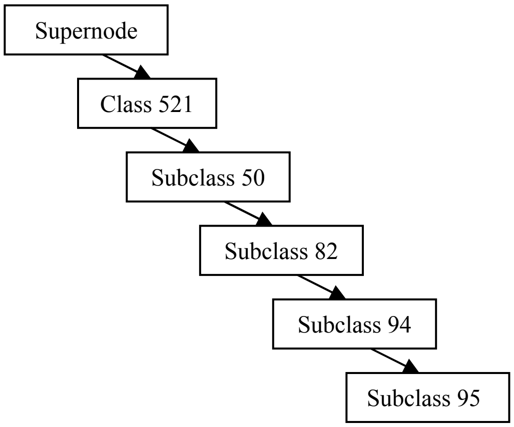

Figure 2. The true USPC Class (521) and subclass hierarchy given for the example in Figure 1

1.2 Hypothesis

As the USPTO classification scheme encodes semantic relationships between classifications, we may be able to use these semantic relationships to rank patents according to the similarity of their classifications. We hypothesize that by ranking patents using these semantic relationships, we can distinguish patents likely to infringe on a certain patent from those which are less likely to infringe.

We present two tasks to test our hypothesis. In both, we use the term *target patent* as the patent for which related patents indicative of prior art violation are to be found from the patent dataset. In IR terms, we may refer to the target patent as the 'query' patent. Likewise, we may refer to the related patents as the 'relevant' patents. The overall goal underlying this hypothesis is to explore the semantic relationship between the classifications in the query patents and those appearing in relevant patents.

Task 1: The premise underlying this task is that the primary classification of the query patent and those of its relevant patents offer meaningful semantic connections. Note that the primary classification consists of a primary class and subclass. We intend to rank candidate patents by estimating how likely they are to infringe on a query patent using a semantic similarity score calculated between the primary classifications of the target and candidate patents.

Task 2: The premise underlying this second task is that the set of *all* classifications of the query patent and those of its relevant patents offer meaningful semantic connections. We intend to rank candidate patents by their relevance estimates using an extension of the method used for task 1. Specifically, we will use the best semantic similarity score calculated between all pairs of classifications of the target and candidate patents. The chemical-related patents we have examined in this study have an average of seven classifications each.

In the above tasks, we view 'semantic connection' as a measurable semantic distance. When two patents have identical classifications their semantic distance is 0, and their semantic similarity is 1. These are elaborated upon in Section 3.1. Additionally, we consider a semantic connection as 'meaningful' if the similarity score computed is one that has a low probability of arising by chance alone. That is, we compare classification similarity scores obtained from related patents with those obtained using random pairings of patents.

Another aspect reflected in the design of our experiments to test this hypothesis is the role of the hierarchy underlying the patent classifications. If the hierarchy is not used the similarity score between two classifications is 1 if they are identical and 0 otherwise. With this baseline strategy, our experiments with all 50 target patents in our dataset yield a mean average precision (MAP) (see Section 3.2) of 0.212 when using the primary classifications only (as in task 1) and 0.265 when all classifications are used (as in task 2). These two baseline runs are represented as Baseline 1 and Baseline 2 respectively. We would like to determine if we can improve upon these MAP scores by considering the hierarchical structure.

The above tasks may seem trivial in the context of the large body of research that exists on classification schemes and their applications. However, as seen in the description provided in Section 1.1, the hierarchical structure underlying the USPC has non-trivial peculiarities. For example, class subdivisions are sometimes motivated less by ontological reasons and more to prevent too many patents from being assigned to a single class. To the best of our knowledge, this is a unique feature in the development and application of a classification scheme. The impact of this 'policy' on using classifications for retrieval is therefore unpredictable. Hence these two tasks are worth further examination.

2. RELATED WORK

Due to the increase in costs of patent infringement litigation and advancements in text mining, a number of techniques have been introduced to examine patent similarity. Initially, much work focused on the categorization of patents into similar groupings. Chakrabarati et al [2] performed small-scale tests using Bayesian classification methods and demonstrated that categorization of patents could be improved by using the classifications of cited patents. Larkey [4] was able to improve the precision of patent searches using the k-Nearest Neighbor approach, but was unable to improve patent categorization.

More recently, attention has been focused on evaluating relevance of candidate patents against a target patent. There are several information retrieval competitions and workshops, such as NCTIR, CLEF and TREC, which focus on patent invalidity searches. Fall et al [3] showed how different measures, when indexed against different sections of a patent's corpus can improve results against the more regimented International Product Code (IPC) structure. However, these competitions and methods focus more on the examination of appropriate query search terms and less on the use of classification to determine patent relevance.

Research in linguistics has focused on evaluating the distance between nodes of hierarchical structures. Shahbaba and Neal [7] have used Bayesian form of multinomial logit (MNL) to improve classification of terms, but this requires prior knowledge of correlations between nodes, which is expensive to calculate. Others such as Leacock and Chodorow [5] and Rada [6] have focused on semantic relatedness of WordNet ontologies. In this paper, we borrow from ontological similarity techniques and extend them to patent classification hierarchies.

3. EXPERIMENTAL DESIGN

3.1 Methodology

We determine the similarity between a target patent and candidate patents by examining the similarity between classifications with respect to a hierarchical classification structure. We begin by processing the true tree structure to represent all USPC classes and subclasses, as shown in Figure 2.

We chose to use a modified version of the Leacock-Chodorow method to produce a similarity measure for USPC patent classifications. Budanitsky and Hirst [1] have shown it performs well relative to other measures in WordNet ontologies and it translates easily to the classification hierarchies of the USPC.

Leacock and Chodorow measure semantic similarity as the negative logarithm of the distance between two nodes a and b, scaled by dividing by twice the depth of the hierarchy:

$$sim_{LC}(a,b) = -\log\left[\frac{dist(a,b)}{2D}\right] \qquad (1)$$

We refine this equation to bound this similarity score between 0 and 1 and modify the similarity equation as follows:

$$sim_{LC'}(a,b) = 1 - \left[\frac{dist(a,b)}{D_a + D_b}\right] \qquad (2)$$

This second equation preserves the ranked order of patents while producing a similarity score in the range of [0, 1], with higher scores representing a better match. Since the resultant similarity score is normalized, it can more easily be used later in combination with scores achieved from other retrieval techniques to improve the precision of a ranked list of candidate patents.

The USPTO has several arbitrarily-divided classes with identical descriptions; for example classes 520-528 all are described by the USPTO as "Synthetic Resins or Natural Rubbers." Likewise, there are 17 separate classes that are described identically by the USPTO as "Organic Compounds." Since it is possible for a candidate patent to be in a different class (and therefore a different subtree) from a target patent and yet invalidate that target patent, we need a method to calculate similarity between classes of disconnected subtrees. To accomplish this, we introduce the concept of a *supernode*. This single supernode represents a parent node to all CAS-identified USPC classes. This supernode allows us to bridge all related classes as a single hierarchical structure.

The denominator of the second equation represents the sum of the maximum depths for each class from the supernode. It is possible that a and b belong to different classes each with different maximum subclass depths. If a and b are both at the maximum depth of their respective class subtrees, their similarity score will be 0; if a and b belong to the same node (same class and subclass), their similarity score will be one[2].

3.2 Experimental Setup

To begin, we establish a hierarchical structure as represented in Figure 2. This was done by processing the nesting of classifications in the USPTO Classification Manual representing this in a tree structure.

Our dataset consists of 50 target patents each having a primary classification determined by CAS to be chemistry-related. For each target patent, we obtain a list of patents that are listed in it as cited references. We refer to this 'relevant' set as the *candidate patents*. We also identify a random pool of 100 chemistry-related patents. For each target query we compute similarity between the target patent and each candidate patent as well as with each randomly selected patent. The manner in which the similarity is calculated between a pair of patents differs for the two tasks.

For the first task, we compute a similarity score between the primary classification of each target patent with the primary classification of each of its candidates and random patents using the method discussed in Section 3.1. For each target patent, we then rank its pool of candidate and random patents by similarity score. We then calculate average precision for the ranking. This is the average of precision scores calculated for each point in the ranking that is held by a relevant candidate patent. These values are averaged across the 50 target queries to yield mean average precision, i.e., MAP. MAP assesses the degree to which relevant (i.e., candidate) patents are ranked above random ones.

[2] In their paper, Leacock and Chodorow count the number of nodes affected, not the number of edges traversed, so the smallest numerator value (in the case where a and b belong to the same class) is 1, not 0. In our study, we count the number of edges between a and b.

For the second task, we use all primary and secondary classifications while calculating similarity scores. If patent *i* and patent *j* have *N* and *M* classifications respectively, then *N* x *M* comparisons are made. The final similarity score between the two is the maximal value calculated across these comparisons. As with task 1, for each target patent we rank its pool of candidate and random patents by their scores and then calculate MAP.

4. RESULTS

In Table 1, we show the comparison between MAP scores of the non-hierarchical baseline and those using our tree structure where only the primary classification for each target and candidate patent were used.

Table 1: MAP and standard deviation for the non-hierarchical baseline and task 1

	MAP	Std. Dev.
Baseline 1	0.212	0.075
Task 1	0.393	0.130

In Table 2, we show the comparison between MAP scores of the non-hierarchical baseline and those using our tree structure where both the primary and secondary classifications for each target and candidate patent were used.

Table 2. MAP and standard deviation for the non-hierarchical baseline and task 2

	MAP	Std. Dev.
Baseline 2	0.265	0.078
Task 2	0.695	0.127

The target patents that produce a high MAP score for task 1 do not necessarily produce a high MAP score for task 2.

Our results show that the method employed in task 2 significantly outperforms the method employed in task 1, based on a two-tailed t-test of the average precision for each query ($p < 0.01$). However, the relatively large standard deviations for both MAP values indicate that the rankings can be somewhat inconsistent. Both methods significantly outperformed their representative baseline ranking methods, which relied on exact matching as opposed to semantic similarity ($p < 0.01$).

5. CONCLUSION

We have shown that it is possible to effectively rank chemistry-related candidate patents based upon their classifications. Ranking candidate patents that may infringe upon a target patent based on the semantic similarity between patents has been shown to be effective. Our results indicate that ranking based on all classes and subclasses is more effective than ranking by primary class and subclass alone.

5.1 Future Work

This work has been an initial investigation on the role of patent classification. We plan to test the effects of a larger patent dataset on MAP, to test using a weight-based supernode to 'penalize' crossing over to different subtrees in our hierarchy, and to observe whether patents in concentrations other than chemistry will produce similar results. Additionally, since the hierarchy structure of the IPC group/subgroup is similar to the USPC class/subclass, we plan to test our similarity measures with the IPC group/subgroup. Finally, given the encouraging results obtained in this study, our next step will be to combine classification with other ranking criteria such as those derived from the abstract and claims fields of the patents.

6. ACKNOWLEDGMENTS

We thank Yelena Mejova, Viet Ha-Thuc, and Natasha Bullock-Rest for their useful comments on this work.

7. REFERENCES

[1] Budanitsky, A. and Hirst, G. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources, Second meeting of the NAACL, Pittsburgh, USA.

[2] Chakrabarti, S., Dom, B., Agrawal, R., and Raghavan, P. 1997. Using Taxonomy, Discriminants, and Signatures for Navigating in Text Databases. In Proceedings of the 23rd international Conference on Very Large Data Bases (August 25 - 29, 1997). M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld, Eds. Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, 446-455.

[3] Fall, C. J., Toresvari, A., Benzineb, K., and Karetka, G. 2003. Automated categorization in the international patent classification. SIGIR Forum, 37 (1), 10-25.

[4] Larkey, L. S. 1999. A patent search and classification system. In Proceedings of the Fourth ACM Conference on Digital Libraries (Berkeley, California, United States, August 11 - 14, 1999). DL '99. ACM, New York, NY, 179-187. DOI= http://doi.acm.org/10.1145/313238.313304

[5] Leacock, C. and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum 1998, 265–283.

[6] Rada, R., Mili, R., Bicknell, E., and Blettner, M. 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems,Man, and Cybernetics, 19(1), 17–30.

[7] Shahbaba, B. and Neal, R. 2007. Improving classification when a class hierarchy is available using a hierarchy-based prior. In Bayesian Analysis. 2(1), 221–238.

[8] United States National Patent Classifications used by CAS. http://www.cas.org/expertise/cascontent/caplus/patcoverage/usnpc.html

[9] United States Patent Office (USPTO). Manual of Patent Examining Procedure (MPEP), July 2008.