Automated Hierarchical Patent Classification (Bachelor's Thesis Proposal)

Leon Weber

December 18, 2015

1 Background

The number of patent applications is growing steadily over the years. In the year 2014, the U.S. Patent and Trademark Office registered 615,243 filed patent applications. In the year 2000, the number was only as high as 315,015 [1]. In order to efficiently process and search those applications, patent documents are annotated with the different areas of technology to which they belong.

This process of annotating the documents can be automated with software systems. Given a patent document and a patent classification scheme, such systems automatically produce suggestions for annotations [2]. Typically, this is achieved by using a machine learning-based classification algorithm trained on a set of annotated patents.

The predominantly used annotation scheme is the International Patent Classification (IPC) [3]. The IPC's structure is hierarchical: It is comprised of a tree of categories with height 13 [4]. For example, a patent could belong to the categories (1) *physical analysis of biological material*, (2) *physical analysis of liquid biological material*, and (3) *physical analysis of blood* [3]. Here, (2) is a subcategory of (1) and (3) is a subcategory of (2). Other patent classification have a similar hierarchical layout [5].

This hierarchical structure poses an interesting challenge for automated patent classification. As opposed to many other classification problems, a classifier would have to take a hierarchy of target classes into account.

2 Goals

In this thesis, we study the problem of hierarchical classification in a specific set of patents. We focus on two goals:

Survey We will provide an extensive survey of methods suitable for the defined task. In particular, we will focus on the fact that there are many possible ways to handle the hierarchy of classes and group the portrayed approaches accordingly.

Evaluation We will choose two to three of the surveyed methods and benchmark those on our data set. The choice of methods will be guided by at least two factors: (1) Promising results on similar tasks and (2) approaches differing from each other.

3 Related Work

3.1 Surveys

There are at least three works that are relevant in the context of this thesis:

[2] provides an extensive survey of hierarchical patent classification. It focuses on the aforementioned classification scheme IPC. It lists results of many methods for tasks in which patent documents have to be annotated with IPC categories. Furthermore, the authors present a detailed framework for grouping approaches of automated hierarchical patent classification.

[6] gives a comprehensive framework for hierarchical classification. It identifies the major differences between known approaches and provides a detailed review of previous work in this field. It is a bit dated and thus misses some more recent approaches like [7].

[8] is an exhaustive survey of approaches to patent classification. It includes a listing of different methods from a bird's eye view and a detailed section on software systems for patent and text classification. The hierarchical nature of the problem is treated as one of many issues to be considered. Obviously, it misses approaches that were published after 2002.

3.2 Competitions

There have been at least two series of competitions including Patent Classification:

The **NTCIR workshops** held in 2005, 2006/07, 2007/08, and 2009/10 had two different tasks including Hierarchical Patent Classification. One was the classification of patent documents with a taxonomy employed by the Japan Patent Office [5] [9]. The other one was to train classifiers for IPC with patent documents and then use those trained classifiers to classify research papers [10] [11]. The results from the most recent workshop [11] indicate that for levels close to the root good results can be achieved (0.8 MAP at subclass level and 0.64 MAP at main group level). At deeper levels, significantly worse results are to be expected (0.45 MAP at subgroup level).

The **CLEF-IP track** included two tasks in which patent documents had to be classified according to IPC [12] [13]. The tasks differed only in the used sub-hierarchy of IPC. Results from [13] provide further support for the hypothesis that the task difficulty increases rapidly with the level depth.

4 Training and Evaluation Corpus

We will evaluate the chosen methods on a corpus of patent documents which are classified according to the Cooperative Patent Classification (CPC) [14]. The CPC is a hierarchical classification scheme whose main part can be represented as a tree. Additionally, there are codes providing extra information which are no nodes of the tree. Those will be disregarded in the context of this thesis.

Because of some properties of the data, using the whole set for evaluation would lead to problems. Thus, we will choose a subset of the data in which the problematic properties are present to a lesser extent.

The features of the data set which we seek to avoid are the following:

There is a huge amount of data. Over 800,000 documents are annotated with over 181,000 distinct categories.

The hierarchy tree has depth 16 and data gets extremely sparse at deeper levels. On average, there are only 8.1 documents per leaf-node category whereas at the fourth level there are 42.8 (without counting documents from lower levels) per node.

The distribution of examples per category is highly skewed. At the fifth level of the hierarchy, there are 39.4 documents per category with a maximum of 11115 and a minimum of 0 (disregarding lower level documents). The standard deviation is 142.5. Other levels show similar characteristics.

As coping with those characteristics lies beyond the scope of this thesis, we will only consider an appropriate subset of the data. We will focus on a sub-tree of the original category tree which will be less broad and less deep than the full tree.

5 Methodology

For the survey, we are going to use the frameworks provided by [2] and [6] to classify the different approaches from the literature. We will cluster similar techniques and focus on problems which are similar to the one we study in this thesis.

For the evaluation, we will split the provided data set into two parts. One will be used for training and cross validation and the other one for the final evaluation of our trained models. We will compare cross validation and final test results for the selected set of methods. While constructing the models, we are going to use as many of the options given by [2] as possible. This will most likely include variations in features, preprocessing, feature selection, feature weighting, and feature extraction cf. [2].

6 Feasibility

We identified three issues which need to be addressed in order to successfully complete the proposed thesis:

Many methods As mentioned before, there are many different methods for Hierarchical Patent Classification. In order to provide an exhaustive survey, we will have to consider many different texts and this will probably be quite time consuming. One solution might be to cluster the methods and then to discuss only representatives of the resulting groups.

Data sparsity We expect data to become very sparse at some level of the class hierarchy. This problem needs to be solved to successfully train complex models. Possible solutions include special methods to construct training data from the hierarchy (cf. [6]) and using simple models which do not require that many data points.

Multi label In IPC, one patent document may belong to several unrelated categories [3]. Evaluation scheme and evaluated methods have to incorporate this.

References

[1] Patent Technology Monitoring Team. *U.S. Patent Statistics Chart Calendar Years 1963 - 2014*. URL: http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us%7B%5
(visited on 11/12/2015).

[2] Juan Carlos Gomez and Marie-francine Moens. "A Survey of Automated Hierarchical Classification of Patents". In: *Professional Search in the Modern World* (2014), pp. 215–249. DOI: 10.1007/978-3-319-12511-4{_}11.

[3] World Intellectual Property Organization. "International Patent Classification". 2015. URL: http://www.wipo.int/export/sites/www/classifications/ipc/en/gui

[4] Daniel Eisinger et al. "Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed." In: *Journal of biomedical semantics* 4 Suppl 1.Suppl 1 (2013), S3. ISSN: 2041-1480. DOI: 10.1186/2041-1480-4-S1-S3. URL: http://www.pubmedcentral.nih.gov/articlere

[5] Makoto Iwayama, Atsushi Fujii, and Noriko Kando. "Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task". In: *Proceedings of NTCIR5 Workshop Meeting* (2005), pp. 366–372. URL: http://research.nii.ac.jp/ntcir/wor

[6] Carlos N. Silla and Alex a. Freitas. "A survey of hierarchical classification across different application domains". In: *Data Mining and Knowledge Discovery* 22.1-2 (2011), pp. 31–72. ISSN: 13845810. DOI: 10.1007/s10618-010-0175-9.

[7] Ricardo Cerri, Rodrigo C. Barros, and André C.P.L.F. de Carvalho. "Hierarchical multi-label classification using local neural networks". In: *Journal of Computer and System Sciences* 80.1 (2014), pp. 39–56. ISSN: 00220000. DOI: 10.1016/j.jcss.2013.03.007. URL: http://linkinghub.elsevier.com/retrie

[8] Cj Fall and K Benzineb. "Literature survey: Issues to be considered in the automatic classification of patents". In: *World Intellectual Property Organization, Oct* (2002), pp. 1–64. URL: http://www.wipo.int/ipc/itos4ipc/ITSupport%7B%5

[9] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. "Overview of the Patent Retrieval Task at the NTCIR-6 Workshop". In: *Search* April 2003 (2007), pp. 359–365. DOI: 10.3115/1119303.1119306. URL: http://research.nii.ac.jp/nt

[10] Hidetsugu Nanba et al. "Overview of the Patent Mining Task at the NTCIR-7 Workshop". In: (2008), pp. 325–332.

[11] Hidetsugu Nanba et al. "Overview of the Patent Mining Task at the NTCIR-8 Workshop". In: *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access* Step 1 (2010), pp. 293–302.

[12] Florina Piroi. "CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain". In: *Information Retrieval* (2010).

[13] Florina Piroi et al. “CLEF-IP 2011: Retrieval in the Intellectual Property Domain”. In: *Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop)* (2011). ISSN: 16130073. URL: `http://www.clef2011.eu/resources/procee`

[14] EPO and USPTO. “Guide to the CPC (Cooperative Patent Classification)”. URL: `http://www.cooperativepatentclassification.org/publications/GuideToTheC`