Comparison of IPC and USPC Classification Systems in Patent Prior Art Searches

Christopher Harris
Informatics Program
The University of Iowa
Iowa City, IA 52242
christopher-harris@uiowa.edu

Robert Arens
Department of Molecular Pathology
Columbia University Medical Center
New York, NY 10032
bob.arens@gmail.com

Padmini Srinivasan
Computer Science Department
and Informatics Program
The University of Iowa
Iowa City, IA 52242
padmini-srinivasan@uiowa.edu

ABSTRACT

Patent classification systems are used to help scrutinize patent applications for possible violations of the novelty and non-obviousness/inventive steps of a patentability test. There are several different patent classification systems in use today, each with a different underlying philosophy and approach. We compare the two most widely-used patent classification systems – the IPC and USPC – and examine their ability to help re-rank patents based on similarity. We observed a significant improvement in MAP, Recall@100, and nDCG when using these systems to re-rank our retrieved document set, demonstrating their overall utility in patent searches.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

1. INTRODUCTION

1.1 Background and Motivation

Patents, which are exclusionary rights issued by the patent office of a national body, are designed to protect the rights of the patent holder, or the patent holder's assignee, for a specific amount of time in exchange for the disclosure of an invention. Due to rapid global technological advancement, patent issuing bodies have witnessed a strong growth in the number of patent applications in recent years – for example, between 1999 and 2009, annual patent applications to the United States Patent and Trademark Office (USPTO) have risen from 288,211 to 482,871, an increase of 67 percent [22]. Other large patent-issuing bodies such as European Patent Office (EPO) and the Japan Patent Office (JPO) have experienced a similar increase in patent applications over the same period.

When a patent application is filed, a patentability test is performed for novelty and for a determination of the non-obviousness nature of the patent application. To perform this task, the patent application is examined against a list of all prior art patents with an earlier priority date[1].

This examination procedure is time-consuming and prone to error. Moreover, the resources available for patent searches are frequently constrained by limitations of time or manpower; hence the need for a ranked list of prior art patents that are similar to the patent application.

Nearly all patent-issuing bodies have some form of patent utility classification. This classification allows patent documents to be easily retrieved and identified [25]. This use of classifications helps to expedite prior art searches, and helps avoid possible ambiguity that may be present in other keyword search fields. In addition, certain patents, such as design patents, have limited text, making keyword searches inefficient. Also for any patent-issuing office to ensure an invention is genuinely novel, it is essential to search those published by patent-issuing offices in other countries. Thus, a harmonized classification system, which is the primary goal of the International Patent Classification (IPC) system, make searches for prior art across the collections of numerous international patent offices more efficient.

There are several major patent classification systems in use today. The World Intellectual Property Office (WIPO) established the IPC, which claims over 71,000 separate classifications down to the subgroup level [5]. Over 100 patent-issuing bodies use the IPC, making it the most widely-used patent classification system. The United States Patent Classification (USPC) system was established by the USPTO, and classifies patents into at least one of approximately 470 classes and 163,000 subclasses. The USPTO uses the USPC primarily, but US-issued patents reference the IPC system as well. Others include the European Classification (ECLA), which has a structure similar to the IPC, and the JPO, which uses multi-dimensional File-Forming Terms (F-Terms) to augment IPC patent classification searches.

Each of these major classification systems was developed with a different underlying philosophy, and this philosophical difference is reflected in the classification system design. In this paper, our objective is to examine whether there is a significant difference in the quality of the ranked list of prior art patents using two different classification systems. To do this, we apply methods that examine the semantic similarity between classification codes

1 In IR terms, patent applications and prior art patents would respectively be referred to as query patents and its appropriate time-sliced dataset. From these we provide a ranking of patents that potentially violate the novelty and non-obviousness/inventive steps of the patentability test.

with the belief that patents with codes similar in a classification code hierarchy are also more similar to each other. We focus on the difference between the IPC and the USPC systems in their ability to retrieve and rank prior art patents. In the remainder of this section we illustrate these systems and discuss the differences between their respective approaches to classification.

1.2 USPC Classification

The biannually-updated USPC classification manual consists of over 163,000 entries, representing each patent function as a single class and subclass. Each patent class and subclass contains a title and description; an explicit relationship is defined between them [20]. Although patent classifications are represented in the patent as class/subclass, these classes and subclasses are organized into a fairly deep hierarchy - it may be necessary to traverse as many as 14 distinct subclass levels to reach a subclass definition (at the bottom of the tree) from its class (at the top of the tree).

For example, US Patent 7,025,942 "Control of lead nitrate addition in gold recovery" indicates Class 423 and Subclass 29 as its primary, or first-listed, classification code. Figure 1 illustrates how this portion of the USPC hierarchy is represented.

CLASS 423, CHEMISTRY OF INORGANIC COMPOUNDS

1 Treating mixture to obtain metal contacting compound
23 . Group IB metal (Cu, Ag, or Au)
27 .. Leaching, washing, or dissolving
29 ... With a cyanide compound

Figure 1. A hierarchical representation of a USPC classification showing nested subclasses

The highest hierarchical level in the USPC is the Class. Here Class 423 is associated with patents involved with "Chemistry of Inorganic Compounds." Subclass 29 is actually four levels deep in the USPC subclass structure under Class 423 [23]. As the number of dots to the left of the subclass description increases, it represents a lower level in the classification hierarchy and a greater level of specificity about the patent's utility. Thus we see that USPC classification codes may appear in rather succinct format in the patent document.

1.3 IPC Classification

The 1971 Strasbourg Agreement established the IPC under WIPO, which divides technology into eight discrete sections. The primary objective of this Agreement was to overcome the difficulties caused by using diverse national patent classification systems.

A patent is assigned to one or more of the 71,000 IPC codes that indicate the related technical field or fields the patent covers. These codes are arranged in a hierarchical, tree-like structure with five distinct components [5]. For example, US Patent 7,025,942 is assigned one IPC code - C22B 11/00. This is illustrated in Figure 2 and Figure 3.

The highest hierarchical level contains the eight sections of the IPC corresponding to very broad technical fields, labeled A through H. For example, Section C involves "Chemistry and Metallurgy". Next, sections are subdivided into classes. The eighth edition of the IPC contains 120 classes. Class C22, for example, involves "Metallurgy; ferrous or non-ferrous alloys; treatment of alloys or non-ferrous metals"

Classes are further subdivided into more than 600 subclasses. Subclass C22B, for example, involves "Production or refining of metals; pretreatment of raw materials." Subclasses are then further divided into main groups and subgroups. Main group symbols end with "/00". Ten percent of all IPC groups are main groups. Group C22B 11/00, a main group, to which US Patent 7,025,942 references, involves "Obtaining noble metals." Like the USPC, the dot notation is used to indicate the number of levels a particular subgroup is from the main group.

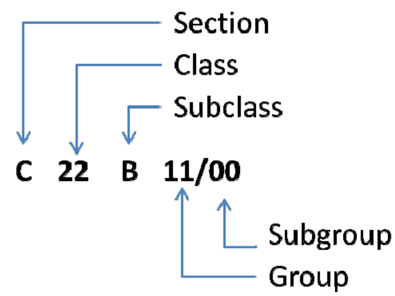

Figure 2. An example illustrating the components of an IPC classification.

C	Chemistry and Metallurgy
C22	Metallurgy; ferrous or non-ferrous alloys; treatment of alloys or non-ferrous metals
C22B	Production or refining of metals; pretreatment of raw materials
C22B 11/00	Obtaining noble metals

Figure 3. A hierarchical representation of the IPC classification provided for C22B 11/00 in Figure 2.

1.4 Differences between the IPC and USPC

With more than double the number of unique classification codes, the USPC has a far tighter focus compared with the IPC. Moreover, in terms of depth of classification, the USPC usually gives more precise information on the invention's true purpose. The IPC classifies an invention according to its function whereas the USPC not only classifies based on the function but also on the industry, anticipated use, intended effect, outcome, and structure. Therefore, the categorization of the USPC is more elaborate than IPC. However, as a result of this narrower focus, the USPC can be more challenging to search. A search across the broader categorization of the IPC system may return a wider variety of patents, improving recall at the expense of precision. Although the USPTO has developed concordance tables to map its USPC system to the appropriate IPC code, these tables are not always accurate nor complete [11]. For example, our examination of these concordance tables for the classification codes listed in US Patent 7,025,942 found that the IPC codes listed as equivalent to USPC code 423/29 do not include the expected IPC code C22B 11/00 [24].

2. RELATED WORK

Although the evaluation of semantic distances in a hierarchy has been examined by numerous researchers in linguistics, particularly, with WordNet, relatively little work has been done in comparing the performance of different patent classification systems. The NTCIR (National Institute of Informatics, Japan) has run several workshops in information retrieval that focused on invalidity searches on Japanese patents [14] but the focus was primarily on classifying documents into IPC or F-Terms. One of the tasks of the 2009 and 2010 TREC-Chemistry tracks sponsored by the NIST (National Institute of Science and Technology) requests participants to produce a ranked list of prior art patents for a set of chemistry-related patent applications [18]. However, the methods used focus primarily on query search terms and have little on the use of classification systems.

The effectiveness of the use of IPC classification system in patent retrieval has had mixed results. In TREC Chem'09, the BiTeM Group discovered that the use of the IPC code actually hindered search quality on chemistry-related patents [6] but did it did not on a broader set of patents from the CLEF-IP 2009 dataset [7]. Earlier research using the IPC at NTCIR found success in its use in comparing patent applications and prior art patents [9][10]. Also at TREC-Chem'09, the Purdue team found that the IPC aided their ability to retrieve prior art patents for a given patent application [2].

Harris et al [8] applied a semantic similarity measure to the USPC code hierarchy, and was able to see a substantial improvement on Mean Average Precision (MAP) over a random selection of patents, but this small study did not compare results from different patent classification systems.

3. EXPERIMENTAL DESIGN

In this section, we develop two hypotheses for examining the role of classification hierarchies in patents and then describe how we established our baseline. We then discuss our methodologies for conducting our experiments.

3.1 Data

We conducted the experiments presented below using patents in XML format from the TREC Chem'09 dataset [18]. Our goal is to compare the quality of the ranked list of prior art patents using the IPC and USPC systems for a given set of 100 patent applications. We constrained the patents we used in this study to a single domain - chemistry. To accomplish this, we focus on those classes determined by CAS (Chemical Abstract Service) to be related to chemistry [3][4], but our methodology is also applicable to other domains.

We created a machine-readable version of the IPC and USPC classifications, ensuring that each IPC and USPC code is represented by its complete hierarchical path; this is discussed further in Section 3.4. This effort was required as each code's full path was not explicitly available in the patent document.

3.2 Hypotheses

Our first hypothesis is that the use of the information contained in the IPC classification hierarchy can aid a patent examiner in finding patents similar to a given patent application. Also we believe that a prior art patent's classification codes need not to be exact matches to those contained in the patent application. However, the closer the classification codes from the patent application and the prior art patent are in proximity with one another in the classification code hierarchy, the more similar the two patents are. Our reasoning is that the classification hierarchy was implicitly designed with this concept of proximity in mind, and this can be useful in retrieving similar patents even if they use different codes.

Our second hypothesis is similar to our first; we examine the ability to re-rank patents based on the USPC classification system based on the semantic similarity of the classification codes between each target and prior art patent, and then evaluate the improvement of the ranking of prior art patents for each patent application.

3.3 Baseline Retrieval and Ranking

We initially retrieved a ranked set of up to 1,000 documents for each of 100 patent applications using Indri [17]. The 100 queries were the first 100 patent applications used in the Prior Art task of TREC-Chem'09 [18]. The retrieval methods used to produce these initial ranked sets did not consider classification codes. Instead they rely on text searches against the patent's title, abstract, claims, and description fields (see the TREC-Chem Run 1 description in [12] for additional information on how this baseline set was retrieved). To make a fair comparison between the two classification systems, our document set was limited to patents with at least one IPC code and one USPC code. In total we obtained a set of 54,364 USPTO-issued patents retrieved by the 100 queries satisfying these criteria.

Of these retrieved patents in our baseline, 2,507 had been judged as relevant patents to the corresponding patent application, using the methods described for the first TREC-Chem run in [12]. Our procedure is to re-rank these prior art patents using our new methods to determine if these methods could be improved by examination of the IPC classification alone or the USPC classification alone. If we are able to significantly improve the ranking of the retrieved patents based solely on the use of one or both of the classification systems, we believe that our methods can make a difference to patent retrieval.

3.4 Methodology

After explicitly deriving and including an XML field representing the hierarchical path to each classification code assigned to the document, we used Indri [17] to index, retrieve, and rank our documents. Although patent documents contain many fields applicable for searches, our focus here is to examine the classification hierarchy's utility in this process, so we focus on the use of the two classification code systems in searches. For example, IPC code C22B 11/08 (a subgroup of the IPC code shown in Figures 2 and 3) is represented in XML as:

```
<ipc-path>
   <class-path> C22B </class-path>
   <group-path> 11:00:11:08 </group-path>
</ipc-path>
```

We added these XML tags to the existing XML for each IPC code appearing in our patent collection; other parts of the IPC code that are not part of the five components mentioned in Section 1.3 were ignored.

The USPC classification hierarchy was also expanded in a similar fashion –adding the XML tags to represent the class and subclass path to the existing XML document. For example, the USPC code shown in Figure 2 is represented in XML as:

```
<uspc-path>
   <uspc-class> 423 </uspc-class>
   <uspc-subcls> 1:23:27:29 </uspc-subcls>
</uspc-path>
```

Documents are then compared for similarity based on their IPC or USPC code fields; we use Indri for this. In particular Indri allows us to compare paths. We note that two codes with similar paths score higher than if they have nothing in common. Here we are storing the full paths as independent tokens. Other experiments evaluating hierarchical structures with Indri demonstrate that if two tokens have more overlap in the included classification codes, they are more closely-related (such as a child, parent, or sibling) in the hierarchy, and they will rank more highly than those which are more distantly-related.

3.5 The Indri Search and Retrieval System

At the core of our system was the Indri search engine, an open source component of the Lemur Language Modeling Toolkit. The retrieval model implemented in Indri combines language modeling [16] with an inference network [15][18][20]. To illustrate, if we have a query q that consists of several query terms (q_1, q_2, ... q_n) and a document d, the occurrence of each of these individual query terms, q_i, are assumed to be independent from the occurrence of the other query terms [26]. Therefore, the likelihood of the entire query can be calculated as the product of the likelihood of each individual query term appearing in a specific document [1]:

$$P\ (q \mid d) = \prod_{q_i} P\ (\ q_i \mid d)$$

Indri allows us to create a separate index for a defined portion of the document (in Indri, the portion of a document is called an *extent*). For example, for a given patent, we could specify a separate extent for each class and for each subclass path, allowing us to combine beliefs, or probabilities of term occurrence, on each extent. This provides substantial flexibility in our retrieval model, including the ability to assign weights to the class extent and the subclass path extent separately. In our experiment, however, we gave the class and subclass path extents identical weights.

The Indri model seeks to determine P(r | θ), or the probability that a particular query term, *r,* occurs in our context language model, θ. In our specific model, for USPC, a patent application and each of our prior art patents are divided into two separate extents: *class* and *subclass*. Smoothing parameters α,β_{class} and $\alpha,\beta_{subclass}$ are applied to each extent respectively. Feature language models θ_{class} and $\theta_{subclass}$ are built specific to each document (patent) in our collection. The IPC follows the same approach, using *class* and *group* instead of *class* and *subclass*, as described in Section 3.4. Indri's inference engine assumes r approximates Bernoulli(θ) [13]:

The retrieval model examines the representation concept nodes, r_i, constructed over our collection model, *C*, based on Bernoulli's conjugate prior, with $\alpha_w = \mu P(w \mid C) + 1$ and $\beta_w = \mu P(\neg w \mid C) + 1$ (Note that μ is a Dirichlet smoothing parameter – for our experiment, μ = 2500). The probability of a representation concept node, r_i, being satisfied by the smoothing parameters α,β_{class} and $\alpha,\beta_{subclass}$ in any given document D is therefore:

$$P\ (r_i \mid \alpha, \beta, D) \quad = \int_{\theta} P\ (r_i \mid \theta)\ P(\theta \mid \alpha, \beta, D)$$

$$= \frac{tf_{w,D} + \mu P(\ w \mid C)}{|D| + \mu}$$

The final step is the creation of the final 'information need' node, which combines the belief node scores into a single score for ranking the prior art patents based on the query terms (classification codes) provided for a given patent application.

4. RESULTS

4.1 Measures

In our experiments, we obtain a ranked sequence of patents from Indri, and therefore it is desirable to consider the order in which the returned patents are presented. The first metric we examine is *MAP (Mean Average Prevision)* as it emphasizes ranking relevant documents higher. Next, we examine *Recall@100*, which is the recall score for the top 100 ranked prior art patents in our retrieval list for each of our patent applications. We use the top 100 as a cut-off; a preliminary examination conducted to determine an appropriate cut-off demonstrated the top 100 to be a sufficient depth. Last, we examine the *Normalized Discounted Cumulative Gain (nDCG)* calculated across the entire set of results for each patent application. Like MAP, this gain is cumulative; however, as we move down our list of results, the gain obtained by finding a relevant document is discounted further.

Table 1 presents a summary of our results. The results for the USPC and IPC were obtained independently, starting from the same baseline. An asterisk (*) next to the number indicates a statistically significantly improvement over our baseline using a two-tailed test with a p<0.05 level of significance.

Table 1. Summary of metrics for our two runs as they compare to our baseline

	Baseline	IPC	USPC
MAP	0.0562	0.0696*	0.0774*
Recall@100	0.2191	0.2802*	0.2912*
nDCG	0.2314	0.2881*	0.3206*

4.2 Influence of Patent Classification Systems

For our first hypothesis, we examined if the application of the IPC classification hierarchy provides a more meaningful set of results than if the classification hierarchy is not considered. Our results show that the IPC classification hierarchy does have an impact; based on our 100 queries. The resulting MAP, Recall @100, and nDCG scores increased by 24%, 28% and 25% respectively. For our second hypothesis, we examined if the application of the USPC classification hierarchy provides a more meaningful set of results compared with our baseline. Results show that the USPC classification hierarchy does have an impact as well. The resulting MAP, Recall@100, and nDCG scores demonstrated increases of 38%, 33% and 39% respectively. Our best results were obtained using the USPC classification system; further analysis using a one-way between-queries ANOVA found that the improvement of the USPC system over the IPC system and baseline was indeed significant at p<0.05 across all three metrics.

5. CONCLUSION

The results from both experiments showed significant improvement across our three metrics over our baseline results for both IPC and USPC, indicating that the use of classification codes has an impact on patent retrieval. This result indicates patents with more closely-related classification codes also are more similar. The USPC system outperformed the IPC system across the three metrics used in our evaluation. This result is not surprising, since the more-granular USPC is primarily used for improving search whereas the IPC is focused on harmonization between patent-issuing bodies; however, no study had measured these results and it was not known how significant this improvement would be before conducting this study.

One limitation of our work is that the initial retrieval set for the 100 queries were constrained to those retrieved by our baseline strategy. It is possible that several relevant documents were excluded by these methods. Our re-ranking efforts are not designed to improve recall and thus inherit the same limitations.

In future research, we plan to examine if the number of classification codes listed in a patent affect the ability to retrieve and rank them accurately; in the patent document set we used for this study, there were an average of 5.7 USPC codes and 3.5 IPC codes per patent. It follows that the greater number of USPC codes provided per patent in our dataset, relative to the number of available IPC codes, may have improved the USPC's ability to discover patents that may contain prior art.

Arguably, the ECLA is a more suitable choice to compare with the USPC as these two classification systems have a similar purpose and a similar level of granularity. The dataset used for our study did not provide data to compare the USPC and ECLA; we plan to compare the quality of results for these two classification systems in a future study.

6. REFERENCES

[1] Cao, G., Nie, J., and Shi, L. NTCIR-7 Patent Mining Experiments at RALI. In the *Proceedings of NTCIR-7 Workshop Meeting*, December 16-19, 2008, pp 347-350.

[2] Cetintas, S. and Luo, S. Strategies for Effective Chemical Information Retrieval. TREC 2009 Notebook. 2009.

[3] Chemical Abstract Service. IPC8 Guaranteed Coverage. http://www.cas.org/expertise/cascontent/caplus/patcoverage/ipcguar8.html

[4] Chemical Abstract Service. USPC Codes used by CAS. http://www.cas.org/expertise/cascontent/caplus/patcoverage/usnpc.html

[5] European Patent Office (EPO). IPC (International Patent Classification). http://www.epo.org/patents/patent-information/ipc-reform.html

[6] Gobeill, J., Teodoro, D., Patscheb, E. and Ruchi P. Report on the TREC 2009 Experiments: Chemical IR Track. TREC 2009 Notebook. 2009.

[7] Gobeill, J., Teodoro, D., Patscheb, E. and Ruchi P. Exploring a wide Range of Simple Pre and Post Processing Strategies for Patent Searching in CLEF IP 2009. CLEF 2009 Working Notes. 2009.

[8] Harris, C. G., Foster, S., Arens, R., and Srinivasan, P. On the Role of Classification in Patent Invalidity Searches. In *Proceedings of the 2nd international Workshop on Patent information Retrieval* (Hong Kong, China, November 06, 2009). PaIR '09. ACM, New York, NY, 29-32.

[9] Itoh, H. 2005. NTCIR-5 Patent Retrieval Experiments at Ricoh. In *Proceedings of NTCIR Workshop 5 Meeting. December 6-9, 2005. Tokyo, Japan.*

[10] Konishi, K. 2005. Query Terms Extraction from Patent Document for Invalidity Search. In *Proceedings of NTCIR Workshop 5 Meeting, December 6-9, 2005, Tokyo, Japan.*

[11] Liu, H. H. Comparison between the IPC and USPC. http://www.ipmall.info/hosted_resources/tools_strategies/bp00/liu.htm

[12] Mejova, Y., Ha Thuc, V., Foster, S., Harris, C., Arens, R., and Srinivasan, P. TREC Blog and TREC Chem: A View from the Corn Fields. TREC 2009 Notebook. 2009.

[13] Metzler, D. Indri Retrieval Model Overview. July 2005. http://ciir.cs.umass.edu/indriretmodel.html. Retrieved on June 22, 2010.

[14] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. Overview of the Patent Mining Task at the NTCIR-8 Workshop. In *Proceedings of the Eighth NTCIR Workshop Meeting*, 2010

[15] Petkova, D. and Croft, W. B. 2007. Proximity-based Document Representation for Named Entity Retrieval. In *Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management*. Lisbon, Portugal, November 06 - 10, 2007. pp 731-740.

[16] Ponte, J. and Croft, B. A Language Modeling Approach to Information Retrieval. In *Proceedings of SIGIR*, pp.275-281, 1998

[17] Si, L., Jin. R., and Callan, J., A Language Modeling Framework for Resource Selection and Results Merging. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, McLean, VA, Nov 4-9, 2002. pp 391-397.

[18] Strohman, T., Metzler, D., Turtle, H., and Croft, B. Indri: A Language-Model Based Search Engine for Complex Queries. In *Proceedings of the International Conference on Intelligence Analysis*. McLean, VA, May 2-6, 2005.

[19] Text Retrieval Conference (TREC). http://trec.nist.gov/.

[20] Turtle, H. and Croft, B. Evaluation of an Inference Network Based Retrieval Model. In *ACM Transactions on Information Systems*, 9(3):187-222, 1991

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

[22] United States Patent and Trademark Office (USPTO) Concordance http://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm

[23] United States Patent Office (USPTO). Manual of Patent Classification (Class 423). http://www.uspto.gov/web/patents/classification/uspc423/sched423.htm

[24] United States Patent Office (USPTO). Concordance Tables for USPC to IPC8 (Class 423, Subclass 29). http://www.uspto.gov/web/patents/classification/uspc423/us423toipc8.htm#C423S029000

[25] World Intellectual Property Office (WIPO). Frequently Asked Questions about the International Patent Classification (IPC). http://www.wipo.int/classifications/ipc/en/faq/index.html

[26] Xu, J. and Croft, B. Query Expansion Using Local and Global Document Analysis. In the *Proceedings of SIGIR*, pp.4-11, 1996.