Patent Mining: A Baseline Approach

Fredric C. Gey[1] and Ray R. Larson[2]
[1]UC Data Archive and [2]School of Information
University of California, Berkeley 94720-5100, USA
gey@berkeley.edu, ray@ischool.berkeley.edu

Abstract

For NTCIR Workshop 7 UC Berkeley participated in both IR4QA and the Patent Mining Tasks. This paper summarizes our approach to Patent Mining. Our focus was upon the US Patent collection, and our methodology was to treat patent mining as an information retrieval task and to aggregate multiple patent classifications from retrieved patent documents. The performance was relatively poor, possibly because of retrieving too many documents, or because of non-utilization of blind feedback techniques.

Keywords: ***NTCIR, Automatic Patent Classification, Patent Mining***

1 Introduction

University of California at Berkeley has participated in all seven NTCIR workshops, concentrating primarily on the Cross-Language Information Retrieval Tasks. In NTCIR-3 we also participated in the Patent Retrieval task [1], thanks to the efforts of our colleague, Aitao Chen, now with Yahoo Research. Our document ranking algorithm, used in all NTCIR workshops so far, is a probability model based using the technique of logistic regression (see Appendix).

The Patent Mining task is based upon the proposition that external resources such as science and engineering documents may be mined for example of 'prior art' which could be used to invalidate a patent claim. A prerequisite to doing this 'mining' is to automatically assign patent classifications to such external publications which usually do not have such classifications. Berkeley participated only in the English patent mining subtask, using the selected English abstracts from the NTCIR 1 and 2 scientific collections.

2 Indexing

The English database for the Patent Mining task consists of using the full text of 990,000 patents from the US Patent Office for the period 1993-2000. The database size is around 33 gigabytes. The following is a fragment of a single patent record with the assigned international patent classification code in bold:

<DOC>
<DOCNO>PATENT-US-GRT-1997-05590429
</DOCNO>
<APP-NO>527986</APP-NO>
<APP-DATE>19950913</APP-DATE>
<PAT-NO>5590429</PAT-NO>
<PAT-TYPE>1</PAT-TYPE>
<PUB-DATE>19970107</PUB-DATE>
<PRI-IPC>A61G 13/00</PRI-IPC>
<IPC-VER>6</IPC-VER>
<PRI-USPC> 5600</PRI-USPC>
<INVENTOR>Boomgaarden; Jonathan C.<tab>Kidd; Harold J.</INVENTOR>
<ASSIGNEE>General Electric Company
</ASSIGNEE>
<TITLE>Electrophysiology table</TITLE>
<ABST>An electrophysiology table includes a dual section table top comprising a pair of elongated planar sections in superimposed planar to planar relationship. One of the table top sections is adapted to be manually moveable axially horizontally over the other section, the other section is electric motor powered for axial horizontal motion. The table is also adapted to be vertically adjusted as well as to be tilted or angulated. </ABST>
<SPEC>DESCRIPTION OF A PREFERRED EMBODIMENT Referring now to FIG. 1, electrophysiology table 10 comprises a fixed supporting base 11 supporting a dual section table top 12 thereon. Support base 11 comprises a first unit 13 containing an electromotive, hydraulic etc. lifting mechanism to vertically position a lift unit 14, as indicated by its oppositely directed arrow, on which … </SPEC>
<CLAIM>What is claimed: 1. 1. An electrophysiology table comprising in combination, (a) a fixed support base, (b) a table top support structure on said base, (c) a dual section table top supported by said table top support structure, (d) a lifting mechanism between said base and said table top support structure to vertically adjust said table top support structure with said dual section table thereon … </CLAIM>

According to the rules of the task, "only <DOC>, <DOCNO>, <TITLE>, <ABST>, <SPEC>, and <CLAIM> fields can be used for categorization purposes." A manual examination of many data records showed that the <SPEC> field was by far the largest and the one which seemed to contain the least semantic content, since it refers to figures and drawings not contained within the text. In testing with the Dry Run topics, we found the processing time to be overwhelming so for the actual runs we discarded the <SPEC> field.

Another obstacle which we did not deal with satisfactorily was the lack of standardized formats for the IPC codes themselves within

the data. For example, both the following codes could be found:

<PRI-IPC>A61B 17/68</PRI-IPC>

<PRI-IPC>A61B 17/068</PRI-IPC>

as well as an occasional leading zero instead of blank separator:

<PRI-IPC>H01M004/50</PRI-IPC>

We created some sed and awk scripts to attempt to normalize the codes into a sortable key, but probably made many mistakes because of these leading zeros. At this point we don't know the effect, if any, these formatting inconsistencies made on our results.

We also discovered afterwards that the trailing "group" classification category /00 as in the example patent above:

<PRI-IPC>A61G 13/00</PRI-IPC>

means the entire group, i.e. all numbers from 01 – maximum. The effect of this on our results is also unknown.

3 Retrieval and aggregation

Berkeley's approach was to use the topic (English title and abstract of NTCIR-1/2 scientific article) as a search query against the patent database, returning the top 1000 ranked patent documents. Because of processing time we did not use blind feedback. As it was each run took more than 24 hours on a high performance Intel-based machine running Linux. From each patent, we extracted the IPC code and sorted and aggregated them to provide a ranked list. The initial aggregation was to count the code occurrences and rank by count. A later aggregation summed up the document probabilities. The third run utilized a classification and clustering method described in the next section

4 Classification/Clustering method

Berkeley's best run (below) utilized a method of classification and clustering developed by Larson, described in [4] and [5]. The basic idea is to combine information extracted from each document with a given classification (we used title information) automatically creating a "pseudo" patent document collection with one document per classification (IPC code) and each containing the titles of all of the the original documents with that classification. Terms in the query (journal abstract) are then used to search these grouped titles using the same logistic regression-based search algorithm as used in other runs and return the IPC codes of the pseudo documents in rank order.

5 Official patent mining results

Berkeley submitted three official Patent Mining runs to the English subtask of the NTCIR-7 Patent Mining task, focusing particularly on our information retrieval approach to classification. Performance of our runs is summarized below and is compared to the maximum performing group (MaxMAP) for the English subtask (the mean average precision (MAP) is expressed as a decimal).

Table 1: Berkeley Official Runs

Run BRKLY	Method /index	Berkeley MAP	MaxMAP
PM-02	Classific. /Clustering	0.1265	0.4886
PM-03	Count of common codes	0.0937	0.4886
PM-04	Sum of doc probabilities	0.0990	0.4886

6 Restricting top retrieved

For the Japanese-only task, the organizers of the Patent Mining task created their own baseline performance run using techniques similar to Berkeley's, only stopping the patent document list arbitrarily at 170 documents per topic [6]. The performance of the organizers' baseline method was substantially better than Berkeley. It would be nice to see their methodology also run for the English-only subtask for comparison. We realized that the organizers C1 system [6] was similar in approach to ours, but restricted the number of patent documents retrieved to 170, instead of the 1000 which we used. We decided to experiment with thresholds of retrieval. The idea is that below some threshold, noise patent documents are retrieved, bringing in noise patent classifications. The following table summarizes a recalculation of method BRKLY-PM-04 using four different thresholds between 100 and 250 documents.

Table 2: Threshold Restriction

Threshold	MAP
100	0.1080
150	0.1090
200	0.1065
250	0.1104
1000†	0.0990

†BRKLY-PM-04 official run

As we can see, restriction to a threshold had a marginal effect on retrieval performance.

7 Fusion of two methods

An examination of the first 10 topics revealed that for the first 5, the runs PM-03 and PM-04 performed substantially better than PM-02, while the reverse was true for topics 6-10. This indicated that a data fusion method might have performed better that any of the three methods. We performed a simple Round Robin merge between the top 100 ranked classifications of methods BRKLY-PM-EN-02 and BRKLY-PM—EN-04. The result, displayed in the next table, shows a substantial gain (44% MAP improvement over BRKLY-PM-EN-02) from the fusion of the two methods, but still a performance substantially below that of other participating groups:

Table 3: Round Robin Merge

Method	MAP
Round Robin Merge	0.1824
BRKLY-PM-EN-02†	0.1265
BRKLY-PM-EN-04†	0.0990

†BRKLY official run

8 Relevance levels in patent mining

An interesting question is whether the methods which perform poorly when for exact patent classification would perform better on a partial classification task. Each of four hierarchical levels of the IPC classification, Class/Subclass/Maingroup/ subgroup could be truncated to higher levels and the effectiveness of the results tested. If at higher levels the recall/precision curves improve, we could consider providing a 'relaxed' relevance for patent classification, as done in the past for the NTCIR monolingual and cross-language IR tasks. Our paper describing this is to be found in the EVIA workshop proceedings [3].

9 Conclusions and future directions

Berkeley participated in NTCIR workshop 7 Patent Mining English-only sub-task by taking a basic information retrieval approach to cross-genre patent classification. Relatively poor performance was observed using these methods. We are proceeding with additional failure analysis of why this occurred.

References

[1] Chen, A and Gey, F. Experiments in Cross-language and Patent Retrieval at NTCIR-3 Workshop, In *Proceedings of the Third NTCIR Workshop on research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo,* 173-182, October 2002.

[2] Cooper W. S., Chen A and Gey F.C. Full Text Retrieval based on Probabilistic Equations with Coefficients Fitted by Logistic Regression. In: Harman DK, ed. *The Second Text Retrieval Conference (TREC-2, NIST Special publication 500-215,* April 1995 pp 57–64.

[3] Gey, F, and Larson, R R, Relevance Levels in *Proceedings of the Second International Workshop on Evaluating Information Access (EVIA-2008),* Tokyo, December 2008.

[4] Larson, R R, "Classification Clustering, Probabilistic Information Retrieval and the Online Catalog." *Library Quarterly*, vol. 61, no. 2 (April), 1991, pp. 133-173.

[5] Larson, R R, Evaluation of advanced retrieval techniques in an experimental online catalog. *Journal of the American Society for Information Science,* 43(1):34–53, 1992.

[6] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto. Overview of the Patent Mining Task at the NTCIR-7 Workshop. *Proceedings of the 7th NTCIR Workshop Meeting on Evlauation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, 2008 (this volume).

Appendix: Document ranking

Berkeley has used a monolingual document ranking algorithm which uses statistical clues found in documents and queries to predict a dichotomous variable (relevance) based upon logistic regression fitting of prior relevance judgments. The exact formula is:

$$\log O(R \mid D,Q) = \log \frac{P(R \mid D,Q)}{1 - P(R \mid D,Q)}$$

$$= \log \frac{P(R \mid D,Q)}{P(\overline{R} \mid D,Q)}$$

$$= -3.51 + 37.4 * x_1 + 0.330 * x_2$$

$$- 0.1937 * x_3 + 0.0929 * x_4$$

where $O(R \mid D,Q)$, $P(R \mid D,Q)$ mean, respectively, *odds* and *probability* of relevance of a document with respect to a query, and

$$x_1 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \frac{qtf_i}{ql + 35}$$

$$x_2 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \log \frac{dtf_i}{dl + 80}$$

$$x_3 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \log \frac{ctf_i}{cl}$$

$$x_4 = n$$

where n is the number of matching terms between a document and a query, and
ql : query length
dl: document length
cl: collection length
qtf_i: the within-query frequency of the ith matching term
dtf_i: the within-document frequency of the ith matching term
ctf_i: the occurrence frequency of the ith matching term in the collection.

This formula has been used since the second TREC conference and for all NTCIR and CLEF cross-language evaluations [2].