WORKSHOP on PATENT RETRIEVAL
SIGIR 2000 Workshop Report

Noriko Kando
National Institute of Informatics (NII)
Tokyo, Japan
Noriko.Kando@nii.ac.jp

Mun-Kew Leong
BIGonthenet.com
Singapore
munkew.leong@BIGonthenet.com

URL: http://www.rd.nacsis.ac.jp/~ntcadm/sigir2000ws/

1 Introduction

The goal of the workshop is to foster research and development of the technology for patent search and retrieval by providing a forum in which researchers and practitioners from relevant communities can share their ideas, approaches, perspectives, and experiences from their work in progress.

Retrieving patent and other Intellectual Property (IP) documents is quite critical in various areas including science and technology, marketing, intellectual property right management, business and so on. In the past, much of the focus for patent and IP retrieval has been from the database community, and not from the Information Retrieval (IR) community. It is partly because the research and development in IR has tended to place emphasis on the generalized systems, which are effective for any kinds of documents and any kinds of queries. It is partly because the text genre of the patents has highly specific characteristics both in semantics and syntax of the text and has highly specialized ways of use by professional users. We assumed that this is the right time to organize a workshop since supporting technology for specialized retrieval is ready.

In addition to the retrieval of documents, other areas of research and development include the extraction, usage and strategic analysis of the text documents, automatic summarization, classification, and comparisons across related documents. A particular area of international interest is the access of patent and IP documents in a foreign language (i.e., what is termed cross-language or translingual document retrieval).

The workshop planed to bring together people interested in patent retrieval and in better access to patent documents from different communities including patent offices, patent vendors, and information retrieval researchers in both system-oriented and user-oriented approaches. The aim was to share ideas, approaches, perspectives, and experiences from their work in progress to ensure a thorough exploration of their common ground. Therefore we placed the emphasis on the following aspects;

- to achieve a better understanding of existing systems, of the user perspective, and of the usage of the patents genre as well as its terminology, ontology, and text structure, and
- to discuss future directions for research and development in patent retrieval and appropriate schemes for evaluation, based on the common ground.

Submitted research papers and invited papers covered wide variety of topics regarding patent retrieval and clearly showed the characteristic aspects of the patent retrieval.

In the following Section 2 shows the format of the workshop. Section 3 is a brief review of the papers presented at the workshop and a discussion, while Section 4 concludes.

2 Format of the Workshop

The workshop started with opening remarks and there were six sessions after it. The first five were about (1) exiting systems for patent searching, (2) research issues in patent retrieval systems including cross language information retrieval, ontology and interaction, and (3) evaluation of patent search systems.

The final session was a discussion with the speakers and audience to discuss issues raised during the workshop and the way ahead of the research and development of the patent retrieval systems, and a demonstration.

3 Workshop Presentations and Discussion

Opening

MunKew Leong, one of the co-chairs of the workshop briefly presented the aim for this workshop and the reason

why we organized it at this SIGIR. Our aims are; to bring together the communities involved, to provide a forum, to foster closer ties between the communities and ultimately to be the first of a series of patent retrieval workshops.

SESSION 1: Existing System for Patent Searching

Munehisa Saito from Japan Patent Information Organization (JAPIO) introduced the commercial patent search system of Japanese patents, utility models, designs, and trademarks PATOLIS. He also discussed on the methods used in its English counterpart PATOLIS-e, which is an information service using the PATOLIS database in English using Japanese-English concordance table and future plan. The presentation gave an overview of the characteristics of the patent search and its difficulties, then also introduced the lexical resources constructed by JAPIO to enhance the access to the Japanese patents.

SESSION 2: Research Issues in Patent Retrieval Systems 1 - Cross Language Information Retrieval (CLIR)

Leo Sarasua and Guido Corremans of European Patent Office presented on "Cross Lingual Issues in Patent Retrieval" and the prototype system ***bSmart,*** which combines phrasal indexing and a Case Based Grammar and has capability of the cross language retrieval among English, German, Spanish, and Japanese. First they summarized the characteristics of patent retrieval and the genre of the patent as (1) the detail of the search, (2) interactive search by experts requires a deep analysis of the expression, (3) structure of a patent document is homogeneous, (4) the heavy use of the generic terms and vague expressions, (5) extensive use of acronyms and new words, (6) sentences are grammatically correct but use limited syntactic structure, which allowing the construction of efficient parsers with very high accuracy. Then the weighting scheme and the results of the experiments of the English-Japanese CLIR was shown. The striking point of the weighting scheme is, the ***tf*** is not used since important concepts are often hidden in the text with general or vague expression in the patent documents.

Masatoshi Fukui, Atsushi Fujii et al, presented on "Applying Hybrid Query Translation Method to Japanese/English Cross-Language Patent Retrieval". It was on the design and results of the experiments on laboratory-based CLIR system with query-translation approach of the patent documents using the algorithm and transliteration, which was originally developed for the Japanese-English CLIR of scientific and technical documents.

SESSION 3: Research Issues in Patent Retrieval Systems 2 - Ontology and Interaction

Ray R. Larson and his colleague presented on "The Entry Vocabulary Module Approach to Patent Search and Classification" and reported the work at the University of California, Berkeley, on the design and development of English language indexes to metadata vocabularies and specifically to the US and International patent classification system. The emerging network environment brings access to an increasing population of heterogeneous repositories with diverse metadata vocabularies. This system is to support searchers who encounter an unfamiliar metadata vocabulary to know which codes or terms will lead to what is wanted.

A.W. McLean presented on "Patent Space Visualisation for Patent Retrieval", which describing a novel and flexible architecture and user interface for patent queries and visualization of query results. Initial sets of requirements were gathered from patent engineers at the company and the author explored the techniques allow a user to build up a stack of queries which allow real time updates of any part of the query and display the results over a 2D map to allow the user to gain a better understanding of the patent space. A query-part can be a normal retrieval query, domain specific (e.g. using International Patent Classification), input, processing, or visualization.

SESSION 4. Research Issues in Patent Retrieval systems 3 - Ontology and Interaction (2)

MunKew Leong presented on "Specialised Querying in Patent Retrieval", which describing a pilot project patent retrieval system between Intellectual Property Office of Singapore. The system is for a specific domain of protein sequence with medium text representation of protein sequences by sequence structure matching of BLAST2.0 and WU-BLAST2.0 homology search algorithms. In addition, other examples of specialized patent search systems, retrieval of chemical patents including chemical structure and/or formula matching and discover chemical reactions, and the process searching were shown. Then the author summarized the characteristics of the patent search in the specialized domain in the aspects of data, search systems, user and workflow and proposed the specialized possibilities.

Naomi Inoue et al presented on "Patent Retrieval System Using Document Filtering Techniques", which describing an experimental patent retrieval system which anybody can make use of easily. It uses document filtering techniques based on probabilistic model searching documents relevant to the user's interest. The system has been used in the authors' laboratories since the beginning of this year and the paper reported the filtering method

used and experimental results.

SESSION 5: Evaluation of Patent Search Systems

Noriko Kando reported on the preliminary results of the brainstorming on "what we shall evaluate on patent retrieval systems" with ten real patent users including specialized intermediaries of patent searching and patent attorneys in the course of planning construction of a patent test collection and an evaluation of patent retrieval systems using it in the future NTCIR Workshop. To set up realistic search requests appropriate for the document types and the real tasks of users is critical to validate the laboratory-typed testing using the test collection. Especially retrieving patents has highly specific characteristics in both the nature of the documents themselves and the way of usage. To understand the real life tasks with patent retrieval is extremely important. The results were summarized in the following points; (1) stages in the patent application process, (2) industries, (3) terminology, (4) high recall, (5) classification scheme, (6) support tools, (7) high precision, (8) document structure (9) images, and (10) task-oriented evaluation.

Preben Hansen and Kalervo Jaervelin presented on "The Information Seeking and Retrieval process at the Swedish Patent Office: Moving from Lab-based to real life work-task environment", which describing a set of methods that is currently used in a study of the task performance process of patent engineers within the Swedish Patent and registration Office and the preliminary results. The focus of the study was to investigate the relationship between the user's work-task and the information seeking and retrieval process. The authors argued that we need to take a broader perspective on the information seeking to understand the task performance process and elicit requirements for information system design.

SESSION 6: Discussion among All the Participants

The final session was chaired by MunKew Leong and was a discussion about the issues emerged among the discussions followed by each presentation. The following issues were raised and discussed with the participants;

(1) Classification schemes,

(2) Tasks -- "information seeking behavior?",

(3) Linguistic and other resources. for example, source data, thesauri, domain word lists, etc. and the importance to encouraging the sharing them,

(4) Language is evolving, unusually fast in the patent context. For example, as Leo clearly pointed out, every patent uses new words. Training based systems must have lag time. and

(5) Interfaces for patent retrieval systems, especially for CLIR.

4 Conclusions

The brief summary of the Workshop on Patent Retrieval was reported. The workshop gathered a wide variety of topics regarding patent retrieval and helped us to understand the characteristics of the patent genre, issues in patent retrieval, user interface and analyzing patents and its users. Also this workshop reported on the current state of the art, expectations from each of the communities and what to expect in the future and provided a good opportunity to learn each other.

For the further information about the workshop, the online proceedings will be available at ;

http://www.rd.nacsis.ac.jp/~ntcadm/sigir2000ws/

It includes the papers presented, slides, and links to the demonstration systems.

ACKNOWLEDGMENTS

Our thanks to the program committee, Mariko Iwasawa, Francois Paradis, and Paraic Sheridan, to the speakers, to the Organizers of the SIGIR 2000 for their help and to all the attendees for their active participation.