Patent Data for IR Research and Evaluation

Mun-Kew Leong, BIGontheNet Pte. Ltd.
munkew.leong@bigonthenet.com

1. Introduction

This is the 2nd NTCIR workshop, and one of its great strengths is that it does not exist in a vacuum. NTCIR owes much to previous workshops, especially to TREC, and currently is one of several complementary workshops for IR and Summarisation research and evaluation. It is therefore both interesting and challenging to look ahead to see how NTCIR can remain true to its cohort yet create an identity of its own and push the state of the art in a research and evaluation workshop. One such way would be to include a new genre of data which is different from what has been done before, which has interesting properties, and which will appeal to sufficiently large number of participants. I suggest that patent data would fit nicely into this role.

2. What's Special about Patent Data?

2.1. Patent data has interesting properties, some of which are listed below:

- 2.1.1. Patent data has structure and semantics, but not in the sense of database records. Here, the documents are primarily text based and are long enough for text-based IR algorithms to apply. The structure of patents is fairly consistent for the genre, with many useful natural classes such as abstract, inventor, assignee, etc.
- 2.1.2. Patent data is very focused. Within the text description, there must be a description of a problem, and of a solution to the problem. In many cases, there is also a method or a process. These all form natural classes as well.
- 2.1.3. Patent data rarely exist in isolation, i.e., it is rare to find only a single patent on some topic. In many cases there are explicit links (in terms of references and citations) from one patent to a prior one. It is also possible to regress the linkages to identify one or more "fundamental patents" which are cited by almost all work in that area.
- 2.1.4. In addition to other related patents, there are also often other technical publications (conference papers, technical reports, etc.) on the same topic.
- 2.1.5. There exists various manually derived hierarchical categories for patents such as the International Patent Classification (IPC). All patents submitted to WIPO (the World Intellectual Property Organisation) are classified into one or more IPC classes.
- 2.1.6. The language and expression in patents is different from that used in political or news-based corpora. It is often highly technical, very directed, and has its own rules and peculiarities. It is also often drafted by lawyers, which may contribute to the characteristics of the language used.
- 2.1.7. Many patents also contain text-like structures such as formulae, equations, sequences, and programming code.
- 2.1.8. Patents are very long documents. I once downloaded the text of the latest 100 patents granted at the USPTO and the average size was almost exactly 100Kbytes.
- 2.1.9. Patents are very often filed in multiple countries, with a high predominance in the USA. In the case where the country of origin is not an English-speaking country, e.g., Japan, then the patent often exists in (at least) two languages; the language or origin and English. Given the commercial value of patents, the patents are often drafted in both languages, i.e., one is rarely merely a translation of the other. This gives parallel corpora which can be used for many purposes.
- 2.1.10. All patents filed with WIPO have abstracts in both English and French. All Japanese patents have abstracts in Japanese and English. This provides a rich corpus of parallel text for investigation, especially as they link back to the full documents.

2.2. Patent data has utility and endurance

2.2.1. Unlike much of the data used in IR evaluation workshops, patent data ages very gracefully. It continues to be searched and referenced long after the patent was filed.
2.2.2. Patents may be thought of as a full disclosure of a method, process, whatever, to the granting body in return for time-limited exclusivity protected by law. Patent data therefore necessarily contains valuable information.
2.2.3. There are systems which provide free search of recent patent data, but charge for searching older patents. This illustrates the fact that patents have both utility (someone is willing to pay for it) and endurance.

2.3. There are alternative methods for searching and accessing patent data
2.3.1. Patents have been around for a long time, and so has searching for them. When patent systems were computerised, so were the searching.
2.3.2. Many non-IR (i.e., non-fulltext) patent search was done on mainframe database systems. Search was restricted to the structured fields in the database records.
2.3.3. More importantly, there are many professional searchers who do nothing but search for patents and patent prior art.

2.4. There are many close areas of research
2.4.1. Other than IR and summarisation, there are other close research fields which have interest in patent data.
2.4.2. These include database retrieval of patents, automatic information extraction (IE) and template filling, automatic abstract generation, translation, etc.

2.5. The IP (intellectual property) business is booming
2.5.1. Last, but not least, patent data is important because there is a global upsurge on the importance of intellectual property
2.5.2. Consequently, there is substantial amounts of funding going into IP search and access.

3. Patent Data will contribute to IR Research

3.1. There is real world demand for better patent retrieval, with users asking for the following:
3.1.1. Better quality retrieval. Everybody wants better precision and better recall. A typical patent assessor (in a patent office) realistically gets about ½ hour to do a prior art search.
3.1.2. Greater transparency. Patent searchers (like many professionally trained searchers) want to know why a particular document is retrieved or rejected. It was relatively easy to understand (even if sometimes incorrectly so) simple Boolean retrieval in a database, and ordering (or sorting) by a given field (e.g., reverse chronological order). It is much more difficult for a lay end user to understand ranking algorithms or pseudo-relevance reweighting schemes. This requires the user to trust the system rather than her or his own searching skills.
3.1.3. Guarantees of completeness. While it's easy for a user to compensate for false positives in the results (just throw them away), it's much more difficult to compensate for missing relevant results.
3.1.4. Multilingual and cross-language searching. This is especially useful for "patent-busting", i.e., when someone is trying to invalidate an existing patent. In particular, to invalidate US patents, patent-busters look to Eastern European patents which are often not in English.

3.2. Scope for additional interesting research
3.2.1. Other than the above, IR is moving towards handling semi-structured data. In particular, IR as a first step followed by IE (information extraction). Efforts such as XML tagging, etc., are steps in that direction.
3.2.2. There is a lot of room for User Interface (UI) research both with respect to the searching of patents, and also to the domain specific applications riding on top of the searches.

4. Opportunities for Summarisation Research with Patent Data

I am not a researcher in the summarisation field, so the points below are possibly suggestions only:

4.1. Natural classes within a single document. As mentioned above, the natural classes within a patent (the problem, solution, etc.) are obvious goals for summarisation engines.

4.2. Natural linkages and operations across multiple documents. There is a demand for text summarisation across multiple documents. The summary may be in text, but may sometimes be more accessible in the form of tables or other structured data. In particular, there is the concept of a "patent map", which is again something done by patent professionals. In a grossly simplified way, it may be thought of as a table of methods against problems. If a

known method has yet to be applied to an existing problem (i.e., that table cell is empty), then there might be an opportunity for research (or a patent) along those lines.

4.3. Built in history and evolution. As mentioned earlier, the maturity of a research topic can be measured by the patents filed in them. This provides many possible targets for summarisation research.

4.4. Fertile area to work across languages. With the large corpora of parallel text in abstract and full patent forms, patent data provide lots of material for research across languages.

5. Taking the Controlled Environment of NTCIR to the Real World

5.1. What's needed to go from a controlled environment to the real world?

5.1.1. Real data with utility, in sufficient completeness for its genre

5.1.2. Provide access to the data using the IR engines to be evaluated

5.1.3. Real users who will use the search engines and whose real world retrieval and use actions can be captured for analysis.

5.2. Patent data and NTCIR can provide the bridge

5.2.1. Patent data, if sufficiently complete (this is flexible depending on alternative avenues of access to the patent information), will form a collection that has utility and which will attract the required users

5.2.2. NTCIR can provide a neutral and anonymous website to provide worldwide access to the data using the IR engines to be evaluated. Since all access will be through the website, NTCIR can measure user actions, etc., through normal web measuring and analysis tools.

5.3. Evaluation of patents in the real world

5.3.1. Moving to the real world abstracts away from "relevance" as the dominating concept in evaluation. There are many problems with defining relevance and maintaining consistency across assessors and between assessors and users.

5.3.2. Evaluation is moved to a user-centric perspective. It is still linked to the concept of relevance but indirectly, i.e., it is assumed that user actions are predicated on the utility of the documents retrieved. Relevance may also be defined as documents of high utility to a given user.

5.3.3. Evaluation in the real world most likely has to be statistically based. This requires a large number of real users, which fortunately is possible if the data has high utility and the interface (website) has a good presence (well designed, fast response, high uptime).

6. A Possible Real World Scenario for Evaluation with Patent Data

6.1. Methodology

6.1.1. Have all the participant systems index the same data

6.1.2. Define a common input and output format, e.g., using XML

6.1.3. Have a website on NTCIR where patent search can be done with the actual search system being anonymous. There can be a common search screen, with results coming from an engine selected at random, or even from all the engines (i.e., metasearch) and displayed individually.

6.1.4. Another possibility would be to make all the engines anonymously available but uniquely identifiable on the website with the same UI for each.

6.2. Metrics

6.2.1. The evaluation must be based on measuring user actions. The performance of the systems is no longer being measured using fixed queries or relevance judgements.

6.2.2. Possible interactions which may be used as metrics include:

6.2.2.1. How the user interacts with the result list

6.2.2.2. Which patent documents are retrieved and from which system

6.3. Analyses

6.3.1. The analysis of the data collected must be statistical, i.e., look for statistically significant patterns of use correlated with particular search systems.

7. Other Opportunities for Research

The setup described in the section above can also be leveraged for research in related areas, including the following:

7.1. Meta-searching. Since each of the search systems are already set up, there is little difficulty in creating an engine that searches all (or some of them) in parallel then merges the results

intelligently before presenting back to the user. It would be possible to use those multiplicity of engines in various ways which may spark off some interesting work in meta-searching.

7.2. Data fusion. This is the instantiation of the “intelligent” merging of results mentioned above.

7.3. Multiple points of entry. Again there is scope for UI research with many systems available through http for comparisons, experiments, etc.

8. Looking Ahead for NTCIR

As I said in the introduction, I believe that by having patent retrieval as a track in the next NTCIR, the workshop will gain in many ways. The patent genre has many interesting challenges and provides a fertile ground for long term research in IR, in summarisation, and in the evaluation of both.