Issues and Non-Issues in Professional Search

John Tait
johntait.net Ltd.
Stockton-on-Tees TS21 3JN, UK
john@johntait.net

Abstract

This position paper points out some false contrasts which are made between Boolean and ranked retrieval, and also between the use in search of statistical machine learning and explicit knowledge representations. Some directions for future research are pointed out.

1. Introduction

This is a position paper on some issues and non-issues in interactive professional search. Specifically I want to talk about some false contrasts which are made, principally between Boolean and ranked searching, and between web searching and professional searching. This paper is based on my experience over the past five years or so of working very closely with patent searchers and more recently with other professional searchers and information analysts.

2. Boolean vs Ranked Searching

Patent searchers and other related professionals searchers (for example the paralegal searchers engaged with the TREC legal track [Tom11]) are often said to prefer Boolean over ranked search systems (like Bing or Google). In fact closer examination reveals they have no problem with the idea of presenting the results which are most likely to be relevant first (the real idea between ranked retrieval). What they are principally looking for is ways to achieve the following:

a. Reproducibility (so they know the same query will produce the same result);
b. An estimate of recall, or at least how likely further work (either on search result analysis or query reformulation) is to reveal additional truly relevant results.

These requirements are the result of the need to potentially defend aspects of the search process. This might be in court: for example was there due diligence in trying to determine whether the launch of new product violates any existing patents; or to a senior manager who has to take business critical decisions on the results of the search.

It is worth noting that reproducibility is a property of the collection as well as the query: hence the IR system really needs to have some way to track former states of the collection, and the changes to those states as well as the query processing. It also means that there is a problem with complex term weighting systems which are often associated with ranked retrieval (especially in web contexts). So, for example, the common technique in web search of effectively reweighting query terms based on breaking news is precisely not what is required for professional search: it produces results which are very time dependent.

It should also be noted that searchers often over estimate the recall they achieve: they are often surprised by results like [Bac11], in which there is clear evidence that there are many potentially relevant documents which Boolean queries cannot retrieve even if hundreds of documents are inspected.

Further it needs to be noted that the notion of relevance is really search task dependent. [Alb11] for instance describes seven kinds of patent search some of which require just a single document to be found (Invalidity

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

Search) and so any further documents found would not be genuinely relevant from the searchers point of view, and others (e.g. Pre-filing patentability search) which require much broader, more familiar notions of topical relevance.

The real point I want to make in this section is that it is often said patent searchers and legal searchers prefer Boolean to ranked retrieval: I believe what they really want is some of the specific properties they obtain from Boolean Retrieval (like reproducibility), and this should be born in mind by researchers and systems vendors.

3. Machine Learning and External Knowledge

One of the reasons modern internet search engines like Google or Bing work so well is that they mine huge amounts of information from the web and from the behavior of the huge numbers of searchers they attract.

But these mining processes rely on sheer scale to get the underlying statistics working in the favor of the search system rather than against it. One of the lessons of the past ten or fifteen years is that techniques, especially statistically based techniques, which fail to work on the small to medium scale will work on the very large scale. Pseudo relevance feedback is often cited as an example, although I'm not entirely convinced by the evidence [Kow00] [Man08]. This is of course because machine learning (and most machine learning is statistical), if presented with a sparse and unrepresentative data set, will tend to learn artifacts of data set and other noise, rather than the underlying true knowledge.

This presents a problem for professional search, and those of us who seek to study and support it. Most professional search is not on sufficient scale to support statistical machine learning, whether in terms of the scale of the document collections, the number of searchers, the number of searches with similar tasks, and so on.

We must therefore look to other routes to include the knowledge obtained through machine learning in our professional search systems.

Note I am not excluding the use of machine learning: rather I am seeking ways to provide alternatives where too little knowledge is available. In might be possible to learn about the 2 million or so independent patents filed annually around the world, but statistical machine learning will not work well on the 44 US tobacco-related patents filed in 2007.

One of the more useful forms of such codified knowledge for search comes from taxonomies and classification. Patent searchers are very used to such search taxonomies, because granted patents are invariably assigned a classification from the International Patent Classification and often another scheme as well. (see [Alb11] for an introductory review, and [Har11] for evidence of the utility of such classification information in search), and recent developments like the Cooperative Patent Classification (see http://www.cooperativepatentclassification.org) and the harmonization activities of the "Big 5 IP Offices" (See http://www.fiveipoffices.org) are only likely to accelerate and extend the usefulness of patent classification as a knowledge source to support search.

Document classification is only one form of semantic knowledge to support search. There are also a number of efforts to build freely accessible standard ontological representations of knowledge in a number of domains – for example in Biomedicine (http://www.obofoundry.org/), or in consumer electronics (http://www.ebusiness-unibw.org/ontologies/consumerelectronics/v1). The new ISO Standard (ISO NP 25964) [Dex11] should also facilitate adoption and interoperability.

Although assessments of the impact of such additional semantic resources on search effectiveness are starting to appear (see [Hua12] and [Bik10] for example) the evidence that they genuinely improve search effectiveness remains sparse. Further in some cases the results may be confounding the impact of topical narrowness of sub-collections with the impact of the use of semantics (see [San12] for a recent study of the impact of sub-collection variation on search effectiveness assessment).

Rigorous studies of the impact of using ontologies and taxonomies in professional search would therefore be a valuable contribution both to the professional searchers and their technology providers.

However, it must be pointed out that there is a problem in integrating machine learned implicit and opaque knowledge with the essentially hand-craft knowledge from the ontologies and taxonomies. They may be different in ways which impact search performance. More particularly the mined information may not reflect the understanding of the expert humans who construct the ontologies. The experts may have knowledge which is simply not available to be learned from the corpus.

Now I am not aware of any studies which reveal this to be a real problem in practice: but then again the relevant studies I have found (like [Hua12] *op cit*) are quite small scale and nothing like refined enough to pick up these sorts of issues.

It must be pointed out that there is a middle ground: semi-automatic or human mediated machine learning, sometimes referred to as active annotation in the Natural Language Processing community (see [Sab12] for a recent relevant survey).

The false contrast I want to point out in this section is the claim that web search is amenable to machine learning because of its scale, whereas professionals search cannot use machine learning because it is too small scale, and therefore must use ontologies and taxonomies. The reality is there are at least two orthogonal dimensions here: scale and accessibility of knowledge. On the scale dimension there may be tiny nuggets of information – Pythagoras' Theorem or the Periodic Table in Chemistry at one extreme and the whole of the web at the other. An on the other dimension, accessibility of the knowledge – the extent to which the knowledge forms part of a codified and agreed body: chemistry versus political sentiment for example.

4. Conclusions

In this paper I have pointed out two false contrasts which are made between general web searching and professional searching. The first false contrast is between the oft-stated preference of professional searchers for Boolean search specifications, and the use of ranked retrieval models usually preferred by IR researchers. The second false contrast is between the opaque and implicit statistical machine learning which underlies much modern web searching and the explicit knowledge representations like ontologies and taxonomies which are often preferred by professional searchers.

For the first, what professional searchers really want is reproducibility of search results, and estimates of recall.

For the second we need to recognize the complex interplay between implicit knowledge mined from the corpus and expert knowledge which may include information which cannot even in principle be obtained by data mining.

Both areas thrown up many opportunities for professional searchers, researchers and technology providers, including assessing the real requirements of various groups of users, providing systems which have appropriate and comprehensible behavior, and which leverage all forms of knowledge to provide an effective search experience.

Acknowledgement

I would like to thank Mihai Lupu for help in the preparation of this paper.

References

[Alb11] D. Alberts, C. Barcelon Yang, D. Fobere-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, & D. DeMarco "Introduction to Patent Searching" in [Lup11], 2011.

[Bac11] R. Bache "Measuring and Improving Access to the Corpus" in [Lup11], 2011.

[Bik10] N. Bikakis, G. Giannopoulos, T. Dalamagas, and T. Sellis. 2010. "Integrating keywords and semantics on document annotation and search". In *Proceedings of the 2010 international conference on On the move to meaningful internet systems: Part II (OTM'10)*, Robert Meersman, Tharam Dillon, and Pilar Herrero (Eds.). Springer-Verlag, Berlin, Heidelberg, 921-938.

[Dex11] S.G. Dextre Clarke "ISO 25964: A standard in support of KOS interoperability" *Proceedings of the ISKO bi-ennial UK Conference* London, 2011. http://www.iskouk.org/conf2011/papers/dextreclarke.pdf

[Har11] C.G. Harris, R. Arens, P. Srinivasan "Using Classification Codes Hierarchies for Patent Prior Art Searches" in [Lup11], 2011.

[Hua12] S.-L. Huang, S.-C. Lin, and Y.-C. Chan. Investigating effectiveness and user acceptance of semantic social tagging for knowledge sharing. *Inf. Process. Manage.* 48, 4 (July 2012), 599-617. DOI=10.1016/j.ipm.2011.07.004 http://dx.doi.org/10.1016/j.ipm.2011.07.004

[Kow00] G.J. Kowalski & M.T. Maybury *Information Storage and Retrieval Systems* 2nd Edition; Kluwer, Norwell, Ma, USA. 2000. p179.

[Lup11] M. Lupu, K. Mayer, J. Tait & A.J. Trippe (eds.) *Current Challenges in Patent Information Retrieval* Springer, 2011.

[Man08] C.D. Manning, P. Raghavan & H. Schütze *Introduction to Information Retrieval* Cambridge University Press, NY, NY, USA. 2008. p171.

[Sab12] M. Sabou, K. Bontcheva, and A. Scharl. 2012. Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW '12)*. ACM, New York, NY, USA, , Article 17 , 8 pages. DOI=10.1145/2362456.2362479 http://doi.acm.org/10.1145/2362456.2362479

[San12] M. Sanderson, A. Turpin, Y. Zhang, and F. Scholer. Differences in effectiveness across sub-collections. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*. ACM, New York, NY, USA, 1965-1969. 2012. DOI=10.1145/2396761.2398553 http://doi.acm.org/10.1145/2396761.2398553

[Tom11] S. Tomlinson & B. Hedlin "Measuring Effectiveness in the TREC Legal Track" in [Lup11], 2011.