The name of the game; information seeking in a professional context.

Jane List,

Extract information,

Cambridge, UK.

jane@extractinfo.info

Abstract

This paper aims to present the needs of the professional information user for data quality. Firstly by looking at quality issues which are important to all professional information seeking tasks, and which should not be overlooked in order for advances in information retrieval to be acceptance by this community; and secondly to provide an example of a specific information retrieval task which, not only requires all the common quality criteria to be met, but is also common to three different professional groups. It is hoped that by considering common tasks such as this, fruitful directions for future research in IR which will benefit a wider number of information seekers will be realised. For the professional information user confidence in quality derives from data integrity, meaning information security, repeatability, transparency, timeliness, and confidentiality of information source, system and solution. Secondly an example of a use case common to three professional user groups will be used to illustrate these quality requirements, and to also describe again the complexity of professional information needs, but also, hopefully to illustrate some further area where information retrieval advances would be welcomed.

1 Introduction

Much research has centred on improving information retrieval for a single domain and for a specific user group, and this has been well received. I think it would be timely, to consider some areas where information retrieval research may focus on the commonalities regardless of domain when considering professional information users and their needs. This approach is not new, Wilson [Wil00]described how, in order to meet the seeking, searching, using information needs, we need to look at sources, and how users can access them, what they do with the information, how they organise the information once it is retrieved, and how this helps users to answer their question, and make a business decision. But in looking again at the common aspects of quality and retrieval across the domains, then this position paper hopes to highlight where acceptance criteria are high and where new approaches for search tool integration and development could be most beneficial.

New and improved ways of meeting these needs will be essential for a positive information retrieval experience and could lead to greater acceptance for new information seeking tools in the workplace. In section 4, a use case is described, whereby the specific information seeking task is related to discovering all of the intellectual property owned by a single organisation. This is actually a non-trivial task, in many cases, and is one which resonates for intellectual property professionals working with patents, designs, trademarks, and in the financial sector. Using this task , examples of where improvements for the IR and search tools could be made.

2   What do we mean by professional search?

Search is now a ubiquitous task performed by all, due to the plethora of information available at any time via a terminal or screen, almost no decision is reached without first performing a search of published information. Mary Meeker [Mee12] gave a thorough overview last year at the NFAIS meeting

But in this paper, I am concerned not with the individual who, in his spare time seeks information for leisure purposes, but the individual who has an information need to fulfil in his professional task. And I propose to outline some essential tasks which require access to information at a deep level, and the characteristics which that information, and the repository in which it is accessed must possess in order to be acceptable to the professional searcher.   The demands made by the professional searcher are high because the consequence of failure (to retrieve the right information) are potentially huge, and could lead to financial losses, have legal or ethical implications or if we take the extreme case of medical or pharmaceutical information error, even lead to patient death.

Professional searchers can be considered as those in University, Hospital and Institute based academic research, researchers in industry, competitive intelligence professionals, and other knowledge workers in government, business and industry and medicine. They all have similar essential needs, as well as needs specific to their domains.   Much research has been carried out , and is still on going, on information seeking behaviour in medical information [Khr11] in the Khresmoi project ; in pharmaceutical information   by the Pistoia Alliance (http://www.pistoiaalliance.org/) and in patent information the IRF initiated some useful research and meetings The IRF , although no longer active, raised awareness with professional users of the possibilities for improvements to search and certainly these three domains are large stake holders, and investors in the search sector.

The fact that professional search is complex, has high demands in terms of retrieval, and is conducted primarily by those who are not trained in information science, means that the job of the database and search tools developers themselves must guide the seeker to the right answer, and even more important in the future will be to make sure they are using the correct sources.   Here I would like to give an overview of some common user requirements, which relate to the provision of data. Addressing these requirements at an early stage is going to be vital to the success of the professional searcher and to developers of professional search tools which must be integrated in to the work flow in an organisation.

3   Common Requirements for Professional Search Solutions

Improving the actual information retrieval function itself is but one aspect of the complete retrieval package. . Professional search tools all offer solutions incorporating these requirements, and the solutions have become embedded in the workflows of the user organisations and are trusted by time-pressed professionals. **(e.g. Thomson Innovation**, **PatBase, Informa.** Common data requirements essential for professional search are outlined below which will influence the success of new technologies and solutions in the professional world.

3.1   Information or IT Security – information and online information system is protected from unauthorised usage, and unauthorised access, and that the information sent as queries and received as solutions are similarly protected. This will be a big issue for forthcoming mobile applications.

3.2   Confidentiality - information seeking and gathering within a business context is almost always subject to confidentiality, particularly so in medical, pharmaceutical and IP fields. It is vital that any online information system, ICT solution or service provided to this community should provide a confidential environment, both in the culture and in the solution. Secrecy in information seeking is important to prevent information falling into competitors' hands, preventing too early disclosure of information to be filed as a patent, that medical records retain patient confidentiality and do not breach ethical guidelines.   For instance query suggestion software commonly used in internet search engines, is usually not acceptable in a professional setting, unless it can be controlled and limited.

3.3   Quality of Information. Information quality is of paramount importance to the success of the players in the professional online information industry.   Quality of Information comprises the following four facets – data integrity, timeliness, repeatability, transparency. These factors are defined, with some examples below:

3.4   Data integrity - ensuring the accuracy and consistency of the data stored in the database, to ensure quality of retrieval.   Data integrity is important at the data creation stage, as well as over the information life-cycle. For instance, patent information derives from the patent specifications published at patent offices around the world. The patent specification is filed according to rules laid down by the patent office, however, for legal reasons there is a great reluctance to change or correct information in the database, even if it is clear that an initial error has never been corrected, e.g. in a name, classification code or other element.   It is then up to the online information providers whether to "add-value" by inserting their own quality control on to this front-line information in order to aid retrieval for their customers. E.g., standard inventor names, and assignee names in some patent databases. Updating processes; and applying changes over the life cycle are also important, but these will not be explored here.

3.5   Timeliness - information must be current. Currency requirements of course vary according to industry, e.g. financial (and some news) information needs to be real time. Patent information is made available in searchable databases on the publication date of the specification by the patent offices themselves; this has driven the professional information providers to up their games in order to compete, even though they must ensure data integrity in their internal systems. It is therefore very important to professional users that the date of availability of the information on the system they use is clear, this also applies to any change date, e.g. of legal status information.   In the better trademark information solutions this information is held at the record and database level, and is transparent to users, this is best practice.

3.6   Repeatability (also called reproducibility) – one of the most important differences between professional and leisure searches is the need for repeatability of the information seeking request, with the need to identify new information and changed information. For instance   Patent searches are very often repeated through-out the life time of the patent to update the information, and to find new related information, to either the document itself e.g. legal status changes, and to additions to the patent family, any changes made to the publication itself, e.g. following grant are also vitally important. IN the Medical / pharmaceutical information field adverse reactions and drug pipeline data also similarly , and vitally, needs to ensure all new and up to date information is available to the searcher and that a repeated search will access the original and the updated information.

3.7   Transparency - this facet when considered from the data and search point of view is paramount to professional searchers, and to intellectual property information searchers, who may need to be able to demonstrate the thoroughness of their research in a legal or business setting, it may be less so to other professionals. However, transparency – can be part of repeatability, data integrity, and therefore quality of search result to the online information industry.

4   Use Case: Name searching - To identify the intellectual property currently owned by a company

To illustrate the requirements I want to consider the search question of identifying the intellectual property (such as patents, trademarks, copyright, designs, trade secrets and know how) held by a particular company. On the surface this seems to be a simple question, but a thorough answer requires some thought. This is a typical research question which must be answered prior to M&A, as part of a licensing investigation, or due diligence activity as part of a patent landscape report, or a competitive intelligence report.

The complete answer will require searching multiple sources, using multiple strategies. The initial step of finding all the company names and individual names which must be searched is outlined below, with an example and a suggestion for a search assistant tool which should be database/source independent:

1) Identify all variations of the company name, including subsidiaries, parent companies, global variations, companies acquired or divested, current and historical names, licensees, and licensors.

Example: 1) A review of patents and trademarks owned by Cambridge Display Technology, would need to incorporate current owner Sumitomo Chemical Company, Next Sierra, Inc, acquired in 2007, Dow Chemicals, Seiko Epson, and Toppan Printing who have partnered with CDT, and in addition for inventors.   This is more important the younger the company, and for US applications e.g.. Prof. Donald Bradley, Prof. Richard Friend, and Dr. Jeremy Burroughes, who all worked at the Cavendish Laboratory of Cambridge University.

Search Integrator: A multi-database name aggregator which automatically selects and searches all name variations, and has a look up facility which is intelligent to consider changing company structure, and searches owner, who-owns-who, applicant, inventor and legal status information. The searcher could select and direct the names on which to search.

2) Analysing the results, using spreadsheets and lists usually takes a lot longer than finding the initial set(s), and visualisation tools could help here.   Of particular importance will be to remove duplicate answers   retrieved   from   the   multiple   sources   searched.

Visual Results Analyser   The results may be needed for different analyses and an image results viewer with multiple views by location, type of IP, subsidiary, lifetime, relationship, legal status would be welcomed.

3) Repeating the search. A search such as this may be updated regularly, particularly in the case of competitive   intelligence.
Search repeater   This tool would automatically update the names, and search the complete set of sources as they are updated, but only deliver to the user the complete report with visualisations, containing only the new pieces of information to a frequency chosen by the professional information user.

5 Conclusion Meeting Professional Users' Needs

This paper has described the professional searcher's concerns over retrieval and data quality, which are different from the casual user of information. The search example, designed to illustrate the quantity of sources which must be searched, and the level of detail and attention required to succeed in a professional task. The three IR tools proposed in section 4 for search, visualisation and repeat tasks, would be good candidates for investigation by the IR and search tools developers. The example was also chosen deliberately to illustrate that searches which are not strictly subject based may also be of interest to the IR community and provide good opportunities for the development and integration of search tools which would be of benefit to users across many industries.

References

[Wil00] Wilson (2000) (http://ptarpp2.uitm.edu.my/ptarpprack/silibus/is772/HumanInfoBehavior.pdf)

[Mee12] - Mary Meeker 12/3/2012 Internet Trends @Stanford – Bases Kick Off