



EDITORIAL

“In theory, practice and theory are the same, but in practice they are not.”

(David Hawking, member of the IRF Scientific Board quoting Lawrence Peter Berra)

The IRF has the mission to bring information retrieval (IR) results to professional information searchers, and to add to IR research this pinch of reality which is required to translate basic research into better products for professional information searchers.

The IRF is a research institute open to scientific and industrial members. Whereas for scientists accessing one of the largest semantic supercomputing infrastructure and large data sets is of high interest, our industrial members appreciate the knowledge transfer from leading IR specialists and the opportunity to outsource R&D projects. The IRF guarantees to deliver a defined outcome in the time and budget agreed. This gives our industrial members the opportunity to free their internal resources and make use of the global science lab.

This first newsletter focuses on the needs of patent searchers and how some current research efforts will contribute to developing better search tools. This is a vast domain to explore, and our next issue will follow up on the challenges of chemical patent searching mentioned in the following pages.

This newsletter will be issued on a quarterly basis and will cover topics which are relevant for information searchers in general and patent searchers in particular. Our aim is to create an additional communication platform between research and industry, to nurture

dialogue and to keep track of research projects with high relevance for the industry.

The IRF newsletter is a service we offer free of charge to our members. If you would like to receive this newsletter in the future, please send an email to membership@ir-facility.org.

Enjoy this first edition of the IRF newsletter!

Your IRF Team

Topics in this 1st issue:

| | |
|---|------------|
| Why we need new approaches to patent searching <i>by Henk Tomas</i> | 2 |
| Semantic annotation: a technology close to market needs <i>by Hamish Cunningham</i> | 3 |
| IP challenges: is there light at the horizon? | 4,5 |
| TREC chemistry track & CLEF-IP: a key to developing better IP search solutions | 6 |
| My IPI-ConfEx 2009 , <i>by Bettina de Jong</i> | 7 |
| Report from the IP Nordic conference | 8 |

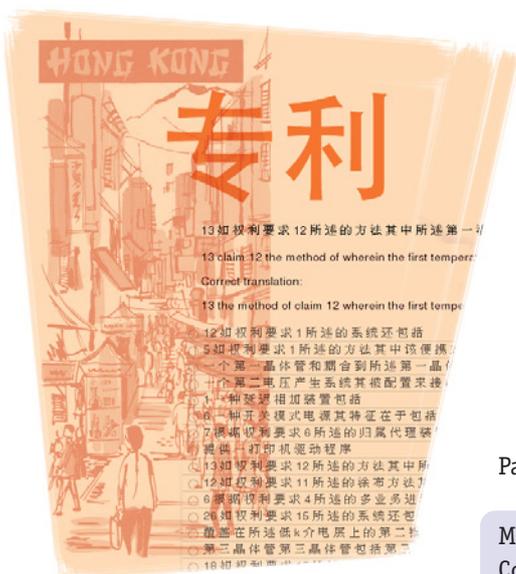
CALL FOR PAPERS /PaIR09: 2nd Workshop on Patent Information Retrieval

PaIR'09 at the CIKM conference in Hong Kong is a workshop which aims to produce the next generation of patent search tools by making patent experts and information retrieval scientists communicate with each other. This year PaIR will have a special emphasis on Asian languages and evaluation of patent search.

The PaIR09 workshop addresses participants from research and industry, who have a strong interest in contributing to the development of breakthrough intellectual property search technology. Patent search experts are invited to present papers on current challenges for the research. They are offered the opportunity to explain their needs to a scientific audience capable of addressing these needs and thus trigger new research projects. The workshop gives them also the possibility to learn which new techniques will be available in the future.

Paper submission deadline: July 17, 2009 | Workshop date: November 6, 2009, Hong Kong

More information on the workshop can be found at: <http://pair.ir-facility.org>
Contact: pair09@ir-facility.org



WHY WE NEED NEW APPROACHES TO PATENT SEARCHING

by *Henk Tomas (Senior Information Specialist, Consultant in IP Information Retrieval)*

If we look back in time a bit, it is only 18 years since we changed from a huge paper patent archive to having the complete text of patent documents on CD.

It is only ten years since IBM (Delphion), Micropatent, the US Patent Office and the European Patent Office started to make patent information available on the Internet.

In 2009 the data of 60 million patent documents is available via the Internet and the complete text of at least 15 million patent documents is searchable, with the numbers growing by at least one million a year.



The challenge of data quality

If we have to deal with such an amount of data, errors will be made by the different parties involved. These parties are the applicants, the patent offices and the database producers.

The patent information specialists want to find all relevant documents in a search. It is beneficial to our community that patent documents are drafted in a clear manner with a descriptive title and abstract and that nobody seeks to obscure the inventive step in the text and claims of the document.

Patent offices could refuse to accept badly structured applications and claims which cannot be searched in every detail. The offices are obliged to check all data and add the proper classification. It is not acceptable that even today IPC classes are missing on some patent documents. Both patent offices and patent database producers invest heavily in correcting patent data, but a lot of errors still remain when patent information specialists are searching in the patent databases. More and more (full-text) patent databases are available and each offers its own added value which differs a little bit from what is offered by the others. Despite all the problems with patent data, more and more patent information is available for free on the internet and for professional patent searchers this information is much more affordable than in the past.

Still a lot of problems remain. Due to globalisation the number of patent applications per year is growing very fast from 1 million filings in 1995 to 1.8 million in 2006. Due to globalisation a lot of documents are in languages which are not common to everybody. More and more machine translations are necessary but unfortunately the quality of these translations is often quite poor. If we look into modern research more interdisciplinary research is done than ever before, which makes it harder to find the relevant documents in a specific technical field.

If we look to the older patent documents, due to restrictions in the Optical Character Recognition (OCR) software, a lot of documents exist with OCR errors and misspellings. Data is missing, and the text and numbers in tables, diagrams, drawings and figures are hardly searchable. If we look to non-patent literature and try to find (prior art) documents on the Internet, it complicates the picture even more.

A long list of unfulfilled wishes

We need new instruments, new features for patent searching. We would like to search in different parts of the documents, not only in the title, abstract, claims and description but also in the examples. We would like to combine chemical compounds with their properties and have the possibility of range searching. We would like to search concepts next to keyword searching. We as professional searchers have a long list of wishes. I am sure that with the help of information retrieval specialists from universities and organisations such as the IRF, next to commercial companies, we can solve some problems. Solutions will be reached much easier and faster, as all parties involved start to cooperate.



SEMANTIC ANNOTATION: A TECHNOLOGY CLOSE TO MARKET NEEDS



An interview with Hamish Cunningham, Research Professor of Internet Computing at the University of Sheffield (UK). Hamish leads the GATE team researching human language computation and is a founding member of the IRF.

During the 2007 and 2008 IRF symposiums, the IP community formulated a number of areas where they have problems when they search in patent (full text) databases. In which areas can you help them with your new semantic annotation tools?

Through our work with Matrixware and Ontotext, we have already begun to show how we can enrich the patent corpus in two tracks: "broad" and "deep" (or vertical and horizontal - language processing works best for simple information, as in our broad track, or very specific information, as in our deep track). For example, some of the annotations we developed last year identify types of measurements found in patents. These are "deep" annotations that selectively identify measurements from other types of numbers. We have also developed "broad" annotations that identify references, such as to figures, literature or other patents. We are now building on this basic set, applying logic on top of the annotations to normalise the measurement units and support searching by range and so on. The ability to search on this information has been directly cited by our colleagues in the IRF as something very desirable.

There is no magic bullet though, and we are not trying to develop artificial intelligences that might replace the skills and experience of human experts. What we are developing are assistive tools and processes that maximise the value of expertise and minimise the costs of applying annotation technology for IP professionals.

What kind of annotations will we see in the near future?

We are currently experimenting with a number of annotation types, including the range searching I just mentioned, as well as identifying very specific types of information that is of interest to various industries. We expect to begin working with biomedical data in patents later this year, for example. More importantly, as we integrate and streamline our processes, we are positioning our technologies to respond quickly to

emerging trends and needs, including those in the commercial market-place.

When can we expect the first results for the end-users in the industry?

We are integrating with the Alexandria¹ database now. It is a big job with more than 70 million complex documents and more than a terabyte of plain text, but we should see some really nice facilities generally available by the end of 2009.

What is the accuracy of the results? Do you think this percentage is high enough for patent search?

There is no simple answer to this question. Certainly, we have been very encouraged by our early accuracy results, but each project and each set of annotations has its own complexities. And accuracy in itself is not a static concept. There are sub-components to accuracy, such as recall (being able to retrieve all possible relevant documents from a search) and precision (the number of relevant documents compared to all documents returned). Accuracy is a blend of those elements, and an acceptable level can depend on what you need rather than raw performance numbers. So that is a long way of saying that sometimes accuracy is in the eye of the beholder. What we are striving for is a system that can be easily modified or customised to meet varying needs. And that is probably what more traditional tools don't do well: adapt.

Do you need input from IP professionals in order to improve your new semantic tools?

Yes, and we can state this very strongly. One of the distinctive characteristics of our approach is that we don't think technology is useful on its own. We can't sell you a revolutionary piece of software that will solve all your problems; what we may be able to do is help you implement robust and sustainable processes that add annotation to your existing tools in order to do some things quicker and easier. In other words, semantic annotation has to be deployed as part of the

workflows of information professionals and with their active engagement and criticism.

What kind of problems do you still have? Is it possible at the moment to annotate text in tables and figures?

Tables: sometimes! Figures and other images: we need to pick up results from other IRF projects that are looking at OCR of image text. But don't expect this to be very accurate in the near future as there are lots of unsolved problems in image processing that aren't going to go away any time soon.

But more generally, the key problems are only partly about annotation itself - we know how to do that pretty well, and as long as the community around the IRF gets interested enough to sustain more work in both the broad and deep tracks we are sure to get more and more useful stuff available (from Alexandria) as time goes on.

What really interests me is making the technology accessible to people without PhDs in language processing - and the problems there are large and will keep us busy for some time!

A lot of companies want to solve their own problems with patent searching. Is semantic annotation a tool for these individual problems?

We are certainly moving in that direction and I think it is one of the most exciting opportunities: to get our tools into the hands of more people who can use them to solve their own particular challenges. That's why we are working to make our tools more interoperable and modular, so they can be put together in a variety of ways. You can see that philosophy in the GATE Teamware suite we deployed for use by the IRF last year. And we have renewed our commitment to training that can support a larger user base, and we are learning from the IRF's commitment to large scale infrastructure and working towards a first release of GATE Cloud. So: lots of exciting stuff to come!

¹ High-quality repository of standardised first level patent literature developed by Matrixware.

IP CHALLENGES: IS THERE LIGHT AT THE HORIZON?

A report from the breakout sessions at the IPI-ConfEx (part 1)

During this year's IPI-ConfEx in Venice, intellectual property experts built up small groups to discuss very specific problems in their domain of expertise.

The outcomes were often presented in the form of wish lists, which gave a clue about industry needs research should try to meet next. In fact, some current research activities might be able to bring first answers.

In this first issue we focus on text mining, chemical structure and biosequence searching. The other discussion topics from IPI-ConfEx will be addressed in the following issues.



Text Mining: a technology that raises high expectations

First a knockdown observation: according to users, no tool currently on the market is solving all problems. Patent searchers at the IPI-ConfEx listed many hurdles, like lack of traceability and flexibility of the systems. Obviously, the general IP framework does not make it easy to develop efficient tools: the accessibility of some full-text sources is limited, additional agreements with publishers are needed.

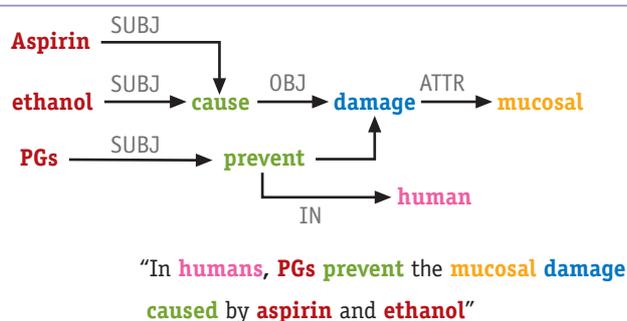
Besides, the amount of unstructured information in patents makes them a difficult source to work with. Patent experts also expressed their concern about the high prices for commercial text mining solutions.

Heard at the IPI-ConfEx:

"Focusing on parts of the text (zoning) is very helpful"



"There is definitely value in analysing unstructured text and not just the indices/codes"



In an ideal, patent-searcher-friendly world, text mining tools would:

- > Allow customisation of thesauri/skill cartridges
- > Be flexible with regard to formats (import & export): The application would work with different databases and allow to merge data from different sources
- > Process all steps (search -> final report) with one tool
- > Deliver different types of analyses according to the very different needs of the customers (e.g. patent attorneys, scientist, marketing)
- > Allow intuitive interpretation of the result, especially if it is interactive

PHASAR: a ray of hope from science

A researcher team around Professor Cornelis Koster at the Radboud University Nijmegen (NL) is currently working on a project called Text Mining for Intellectual Property. They intend to develop a professional search engine based on deep linguistic techniques (PHASAR) as well as an accurate parser for complicated technical English texts (AEGIR). By combining both, they will create a Text Mining system for IP search, which will offer sentence co-occurrence of terms additionally to the current state-of-the-art document co-occurrence of terms.

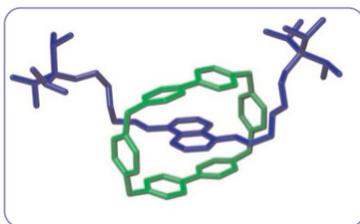
Which benefits will PHASAR bring to the IP community?

- > Ability to work on full-text documents, patent applications, journal article, dissertations and even the whole internet.
- > Support for analysis: classification techniques for document selection and presentation, search within search, interactive construction of reusable search profiles, aggregation of the information from different documents
- > Exact match with full transparency
- > Explicit mechanisms for control over precision and recall
- > Qualitative and quantitative feedback from index and thesaurus

Parser and search engines prototypes are expected to be developed by the end of 2009. The final versions should be delivered by the end of 2011. More information at: http://www.ir-facility.org/the_irf/current-projects/tmip/

Chemical Structure & Biosequence Searching:

Chemical experts at the IPI-ConfEx were well aware of the shortcomings of indexing and retrieval systems: most of them were released in the 80's and were neither



designed to cope with an ever increasing amount of documents nor with the complexity of chemical patents. Usually, out of 1.000 hits, only 5 to 6 are relevant. It is common practice to apply various search strategies to one search problem to generate reasonable results, e.g. permutate by ATOM and by CLASS searches.

Another well known problem is that patent attorneys invent funny, non-scientific names for structures, so that they can never be found; a practice which should not be admitted by patent offices.

It is considered paramount by the patent experts to use multiple databases and tools to improve the quality of the search results. There are different types of searches: macromolecular structure, reactions, processes and/or conditions (e.g. temperature, pressure, etc). Patent searchers complain about the difficulty to find documents older than 20 years, although these documents are important.

The chemical patents experts at the IPI-ConfEx have issued the following wish list:

- > Restrictions on claims in Markush (e.g. no nested R-groups)
- > Better Markush indexing
- > Removal of 99 substance limit in WPI
- > Prophetics back file
- > Easily find location of substance in the document (page, paragraph, table...)
- > Analysis of "gap" between examples & Markush structure to facilitate challenge

Such challenges require the involvement of a larger community of academics, as well as a standardised method of evaluating chemical retrieval tools. For this reason, the IRF is organising the TREC Chemistry track: to present the specific problems of chemistry retrieval to information retrieval academics and to generate, in the longer term, a standard evaluation method for the diverse set of tools which are already existing.

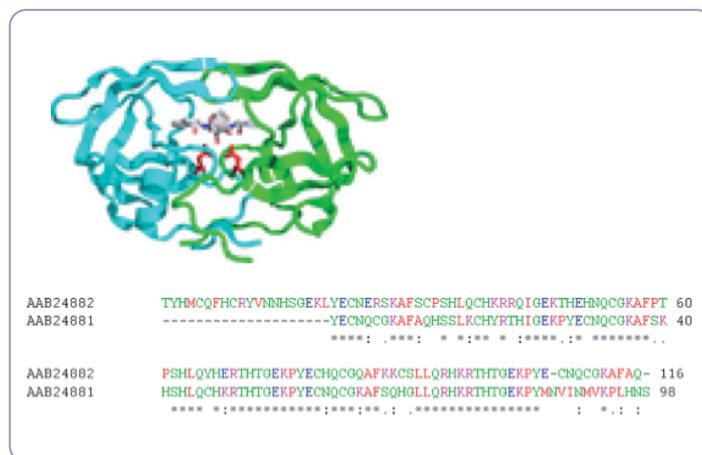
Another group of experts at the IPI-ConfEx has dealt with the challenges of biosequence searching. They observed that none of the sequence databases is complete, so all must be searched. They most widely used databases are STN, GenomeQuest, NCBI and EBI. NCBI uses the BLAST algorithm to find sequence similarities, while EBI provides FASTA.

A good strategy for BLAST/FASTA search is the following:

1. Turn off all filters
2. Reduce results by setting E-filters
3. When searching with decreasing sequence length
 - ▲ BLOSUM value
 - ▼ PAM value
 - ▼ wordsize
 - ▲ E-filter

It is an interesting fact that sequences can be retrieved from public databases, which are not found in commercial databases. As for sequence search tools, the ultimate solution would be a tool which is able to search the available databases in one run and deliver a manageable output.

Furthermore, patent searchers recommend a supplementary text search in addition to the sequence similarity search. It is important to include synonyms, aliases and symbols in your search.



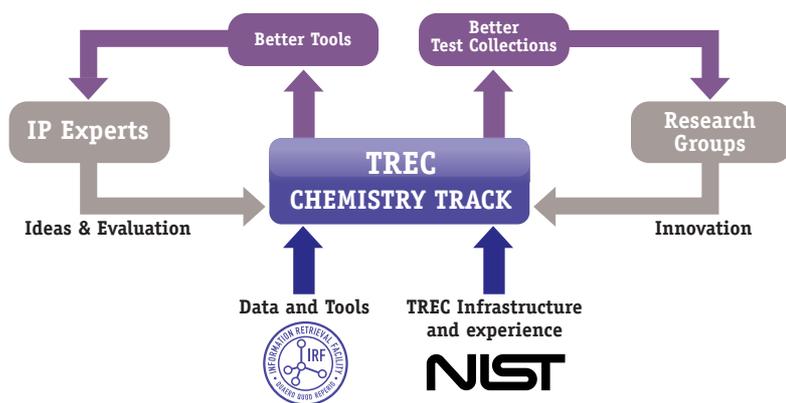
So much about the challenges chemical patent searchers are facing. But what is research undertaking to solve these problems? Are there promising new technologies? We will try to answer these questions in our next issue of the IRF newsletter.

CALL FOR PARTICIPATION:

The IRF plans to go more into the depths of what current research can do to answer the challenges of chemical structure and biosequence searching. We are looking for patent experts willing to share their particular business case with our scientific team and by doing so, giving inspiration for new research projects.

If you are a chemical patent searcher interested in contributing to the development of a technology that is able to solve some of your daily search problems, please contact us at: chemsearch@ir-facility.org.

TREC CHEMISTRY TRACK



The Chemistry Track at the Text Retrieval Conference in the USA, organised by the IRF in collaboration with the University College London and York University, aims to develop methods to evaluate chemical retrieval systems.

Researching evaluation methods may seem less interesting than researching actual retrieval systems, or chemical entity extraction systems, but it is a sine qua non component of the entire information access field. Without the proper means to evaluate retrieval systems, we would rely on hear-say, marketing campaigns and eye-

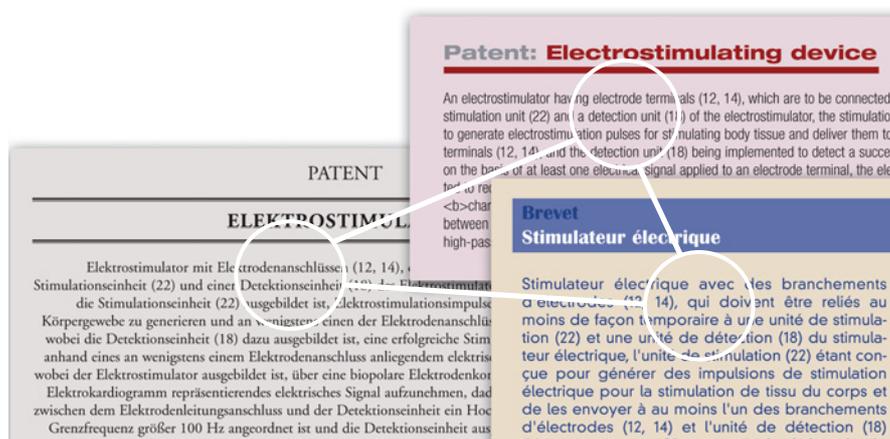
catching graphical user interfaces, rather than the real power of the engine to deliver the results we expect.

This is the first such campaign that focuses on the chemical domain. The final aim of this 3 year project is to provide the industry, as well as researchers, with reliable methods to judge the quality of a chemical retrieval system. The main aim in the first year to establish proper communications channels between the industry and the researchers participating in the evaluation.

The role of the IRF in this track, apart from organisational issues common to any collaborative effort, is to be the bridge between patent experts and information retrieval researchers. Communication between the IR and IP groups has proven to be difficult due to vast differences in terminologies. A common lexicon as basis for mutual understanding needs to be established, and the IRF will collect questions and answers from either side, for ever more efficient tracks and collaboration in the future.

Further information about TREC-CHEM is available at:
http://www.ir-facility.org/the_irf/trec_chem.htm

CLEF-IP: A KEY TO DEVELOPING BETTER IP SEARCH SOLUTIONS



CLEF-IP is an evaluation campaign initiated this year by the IRF within the cross language evaluation forum, a European initiative for the promotion of R&D in multilingual information access. The rationale of an evaluation campaign is to conduct experiments aimed at obtaining reliable measures for the effectiveness of systems and technologies in a specific area of interest. These measures allow researchers to assess, compare, and ultimately improve different approaches.

In its first year, CLEF-IP is going to tackle the problem of prior-art search, one of the the main tasks of a patent examiner. The track will focus in particular on the comparison of different retrieval approaches, the study of automatic query formulation

and multilingual issues. A concrete example of an experiment that track participants are going to run with patent data is: deploying machine learning techniques with a training set built by collecting prior art citations from patents and their families.

One desirable side-effect of the CLEF-IP campaign is a closer cooperation and communication between information retrieval scientists and IP search professionals. Since the latter are the ultimate judges of the quality of search results, a big part of the challenge of our campaign is to

model our measures according to their concrete information needs. And this deep understanding can only be achieved through a constant dialogue and a common language that allows us to see where the goals of the IP and IR communities converge.

The CLEF-IP track is supported by the IRF in the belief that only the assessment and in-depth analysis of existing search technologies make it possible to identify those features that are suitable for improvement and that can yield innovative products.

Further information about CLEF-IP is available at:
www.ir-facility.org/the_irf/clef-ip09-track

MY IPI-CONFEX 2009 by Bettina de Jong (Shell International B.V.)



The IPI-ConfEx 2009 was held from 1 to 4 March in Venice-Mestre, Italy, an appropriate business environment to enable participants to focus on the conference.

A theme throughout the conference was on one hand the need for further development in tools to search and analyse patent information, on the other hand the actual developments in this field. The commercial providers showed further refinements with a range of visualisation tools to analyse search results, relationships, citations etc.

However, especially the non-commercial organisations such as research institutes are taking steps on new paths to, e.g., patent valuation and patent image retrieval. There is still a long way to go in these areas, but there is a clear need from users to be given tools to efficiently manage the ever increasing amount of patent publications. In this respect, the IRF was mentioned several times as a platform that gives special attention to these needs and facilitates the development of solutions. It was a good idea of the IRF to bring some young men from universities who enthusiastically told the conference participants about the projects they are working on.

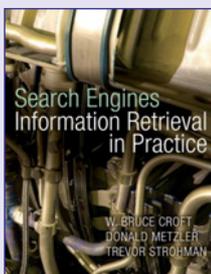
Networking and the exchange of experience being the major aspects of this conference, the round table discussions were again a big success and resulted in several interesting conclusions and recommendations. It would be good to see a follow up on these recommendations, for instance via further discussions or presentations at next year's conference, or via some kind of fora that remain active in between conferences.

One of the other highlights of the conference, the IPI MasterClass™ by Stephen Adams, explained the importance of non-patent literature for the patent searcher. Stephen showed that despite developments in tools and systems, users still need to have a thorough knowledge on what is in the systems and how to use them.

Patent search and analysis remains a complex task, and the IPI-ConfEx, along with showing how the systems are developing, also helps in the development of the patent information specialists.

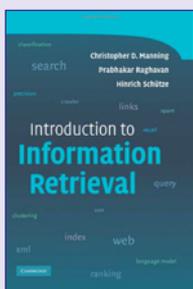


Literature Tips:



Search Engines: Information Retrieval in Practice,
by Bruce Croft, Donald Metzler, Trevor Strohman (Addison-Wesley, 2009)

Written by a leader in the field of information retrieval, this text provides the background and tools needed to evaluate, compare and modify search engines. Coverage of the underlying IR and mathematical models reinforce key concepts. Numerous programming exercises make extensive use of Galago, a Java-based open source search engine. A valuable tool for search engine and information retrieval professionals.



Introduction to Information Retrieval
by Chris Manning, Prabhakar Raghavan and Hinrich Schuetze (Cambridge University Press, 2008)

"This is the first book that gives you a complete picture of the complications that arise in building a modern web-scale search engine. You'll learn about ranking SVMs, XML, DNS, and LSI. You'll discover the seedy underworld of spam, cloaking, and doorway pages. You'll see how MapReduce and other approaches to parallelism allow us to go beyond megabytes and to efficiently manage petabytes."

Peter Norvig, Director of Research, Google Inc.

REPORT FROM THE NORDIC IP CONFERENCE

From 24 to 25 March 2009, patent portfolio managers and patent lawyers convened in Stockholm to share their views on how to best capitalise on inventions.



The full range of possible licensing strategies was presented at the conference: from in-house innovation and implementation without any outlicensing, as it was discussed to be the case for the steel industry, to blended models which are applied by companies like Shell.

In such blended models, innovation is being done in-house, but the implementation of such innovative products that belong to auxiliary lines of business (e.g. drilling technology for oil companies) would be done with the support of venture capital by third parties.

Companies like Nokia or Ericsson do substantial in-house R&D, but also fully offer their technology via licensing to third parties. Other corporations like France Telecom gave insights into their open innovation strategies which are based on the belief that the

best way to access the distributed knowledge of innovative people is by building a network of open innovation.

The enforcement of patent rights in Asia was another special focus area. Interesting insights were given from Chinese patent attorneys with successful examples of patent law suits from European companies in China. It also became clear that dealing with Chinese or Indian patent offices is still difficult for Western companies.

In general, there was a strong agreement that, if the Asian market is within your target market, an intimate knowledge of prior art patents in your respective field, and local professionals that allow you to deal with Asian patent administrations in their local language, are the keys to success.

GET YOUR OWN RESEARCH PROJECT STARTED WITH THE IRF

The IRF not only offers a wide portfolio of standard services, it can also help you find the right scientific partner to develop a tailor-made solution to your particular search problem.

The IRF has among its members leading universities, institutes and students that are committed to do applied IR research, and have passed the high scientific standards of being admitted as IRF members.

This brain power is at your disposal. By submitting your research interests to us, you will be offered a dedicated research project with defined outcome and deliverables. The IRF not only gives you access to outstanding scientific IR expertise, it also allows you to free your own personal resources.

The IRF is equally of interest to you if you are looking for highly qualified, motivated, young IR professionals. We hold a pool of graduates that could be the right match for your vacancies.

For more information, please contact:
Petra Wolf, IRF Membership Services
membership@ir-facility.org



IRF members access for free:

- > the quarterly newsletter with the latest news on applied information retrieval
- > a web-based IP/IR forum that gives access to the knowledge of leading experts from science and industry

Additional services subject to an annual fee include e.g.:

- > in-house workshops with leading IR experts
- > free access to webinars on IP related topics
- > reduced registration fees for the IRF Symposium

Contact us for more information!

Did our first newsletter meet your expectations?
Which topics would you like to read about in the next issues?
Please send your comments and suggestions to
newsletter@ir-facility.org

Imprint:

Information Retrieval Facility Society

Eschenbachgasse 11 | A-1010 Vienna | Austria

Phone: +43-1-236 94 74 | Fax +43-1-585 01 41 | www.ir-facility.org