



## Order out of Chaos the Xapian solution chosen by leading law firm

Mills & Reeve is one of the UK's leading full-service law firms. Like many similar organisations, the company has an archive comprising millions of documents containing everything from emails to letters to past cases, held in a number of different computer systems. However it has become increasingly difficult for staff to find information effectively, as each system has a different user interface and its own particular retrieval 'quirks'.

To address this the firm has turned to the open source search engine Xapian, and has engaged Lemur Consulting to investigate how best to improve its information retrieval systems. Lemur used questionnaires and interviewed staff as well as carrying out technical investigations into the existing systems. Lemur

then designed a new system that can read all Mills & Reeve's existing data into a searchable index, using Xapian. A simple but powerful search interface will allow staff to search all, or some, of the archive and at a glance see if relevant information exists.

Charlie Hull of Lemur said "Many employees at Mills & Reeve were asking for something like Google – a simple, clean interface and a clear list of relevant results. However, for this kind of data the Google engine isn't necessarily the right answer – so we've taken the best of Google – a simple interface – and combined this with an advanced probabilistic search engine, Xapian, to design a powerful, accurate

and expandable system."

Stuart Langridge, Information Architect at Mills & Reeve, said "I went looking for a search engine that was open-source, because open-source technologies are generally



Mills and Reeve Cambridge Office

more robust, easier to use, cheaper, and not driven by the desire to restrict customers. I chose Lemur Consulting because Xapian seemed like the right technology: the Lemur team has extensive knowledge of that technology and reputable names in its client list."

## What is Bayesian Search?

All search engines work in roughly the same way. Starting with a collection of documents, they build an index of terms - words or phrases occurring in documents. This is analogous to an index in a book, and allows documents (or in this case, pages) to be found which contain a particular term. In practice, a document might be a web page, a legal case note, a customer record, etc.

When a user searches for one or more terms, the search engine returns a list of matching documents. The important thing is to ensure the most relevant documents are at the top of the list, immediately accessible to the user. This is particularly crucial for large collections of documents, when there may be thousands of results returned. Many search engines employ simplistic methods such as counting the number of terms on a page, which perform badly and fail the user.

Probabilistic search engines were invented to solve this problem. Based on the Bayesian model of probability, they examine the distribution of terms in the entire collection of documents, and can thus judge how important a particular term is in a query and a document. When this is applied to each document which matches a query, a ranked list of results is produced with the most relevant documents at the top.

The mathematical justification for this approach has been known since the 1970s; the challenge in computing terms is implementing this approach in a fast and efficient way for collections of millions or even billions of documents. Lemur specialises in developing new Bayesian systems and integrating Open Source Bayesian search engines for clients.

## ECIR 2006

<http://ecir2006.soi.city.ac.uk/>

Lemur is pleased to sponsor the ECIR 2006 conference in London. ECIR is the foremost European conference for new research in information retrieval, and has an illustrious 28 year history. As search technologies have become increasingly widespread, the importance of building robust and effective systems, and comparing the performance of these objectively, has become ever more apparent. Many of the latest technologies discussed at earlier ECIR conferences have been incorporated into products developed by Lemur. We are pleased to support the academic community by sponsoring ECIR this year at Imperial College London 10-12 April 2006.

### About Lemur

Lemur offers expert systems advice and customised information retrieval software. As search engine specialists, we can:

- Provide your website or intranet with a high performance search engine
- Offer multimedia database solutions with probabilistic search and XML editing facilities
- Produce information retrieval software for specific purposes; such as building tools that will display data in graphical form, or sort documents automatically.

Lemur also provides general expertise and development in technologies such as Java, C++, Linux, Windows and XML. "We promise to explain ourselves in a clear and jargon free manner, giving you accurate and honest advice."