Legal Information Management, 7 (2007), pp. 258–262
 Printed in the United Kingdom doi:10.1017/S1472669607002071

Searching for the Answer 2.0

Abstract: This article by Paul Billingham of Concept Searching and Phil Ayton of UC Logic covers the history of computerised searching aids in the context of knowledge management projects and describes conceptSearching, which is a scientific approach to trying to solve natural language searching problems using a number of techniques, not just an algorithm. They also consider federated searching and describe the Sysero information portal and a taxonomy manager application.
Keywords: enterprise information management; taxonomies; federated search; integrated search

Introduction

We appear to have moved over to a two point zero world with Web 2.0, library 2.0 and even KM 2.0 being touted in publications and trade shows up and down the country. Put in a nutshell, Web 2.0 is all about ways to publish information, be it documents, web pages, blogs, wikis or video. In this article we will be looking at one area of technology a Knowledge Manager 2.0 can use to get the relevant information to their firms.

Search and classification technology

Around the year 2000, many law firms invested heavily in search and classification technology. Yahoo was still the market leader for internet search and much was made of their hand-crafted taxonomy, where they would manually examine every website and add it to the appropriate node. Within 12 months Yahoo lost their domination to free-text specialists Google, who started making headway with their now legendary PageRank™ system. The exact algorithms behind PageRank™ are a closely guarded secret and change continuously, but the basic principle is that web pages with the most links to them get the top positions in a search. Whilst peer review dictates where a site appears in the hit list, Google is basically an extremely high capacity Boolean search engine.

Document ranking and within document frequency analysis

Traditional free text systems are based on simple keywords and Boolean logic (primarily the AND, OR and NOT operators). Whilst this technique is very precise, it does fall down when the number of documents retrieved is too large to examine exhaustively. In this case the ability to rank documents, with the most important ones at the top of the list, is of paramount importance. Over time traditional systems have introduced various ways to rank results, such as PageRank™, but ultimately Boolean search engines are not based on a sophisticated model of term profiles across the collection of indexed documents and tend to rely on a within document frequency (*wdf*) analysis.

It is fair to say that virtually all search engines are based on *wdf* analysis to some extent, with Bayesian Analysis being one of the most popular approaches. Whilst Bayesian Analysis will find documents that match the words in a user's search, the ranking of these will often be of poor quality as words in isolation, unless they are very unusual technical terms, are fairly meaningless. Therefore the more sophisticated search engines, such as those used for research, add additional techniques for ranking the documents returned using *wdf* techniques. Whilst PageRank™ works well on the internet, documents are generally not connected by hard coded links or references and the process breaks down when indexing document repositories and libraries.

Conceptual queries and document themes - conceptSearching

If you ask for information about a tank, did you mean oxygen tank, Chieftain tank, or septic tank? It would be handy to have a search service that understands the context in which a user is working, and there have been systems capable of doing this for some time. However, the maintenance overhead of such systems tends to be prohibitive and as context, especially in the legal world,

changes with time, the original training processes can quickly become out-of-date and therefore irrelevant. conceptSearching indexes are based on an on-going statistical profile of the corpus of documents which creates the effect of a self-learning and automatically updating index which reflects new terminology, acronyms and dynamic changes based on the index content.

The advent of conceptSearching indexing would appear to have the benefit of good political timing as well as technical advantages. If systems that require user training are unpopular, the second most obvious solution for providing more accurate searching would be the type of personal search systems that Google recently got itself in much hot water over. In this case it's the system that generates a statistical profile of the user and directs his further searches accordingly. However, the US does not have an equivalent to our Data Protection Act and so, even if a statistical profile of user search habits did provide greater search accuracy, its use within the EU would be much restricted. The big brother implications I will leave to the tabloids.

With the above in mind, it would appear that any search system that tries to extract information from the user is either high maintenance or legally and morally perilous. Therefore, any advanced search capability must come from within the information being researched, rather than from the user's environment.

Relevancy ranking based on conceptual understanding

The move from Boolean-based search to conceptSearch has had a few dead ends in the past. Artificial intelligence and natural language techniques dating back to the 60's have spectacularly failed to deliver, with the Turing Test still running 50 years on. conceptSearching is a more scientific approach to the natural language search problem and is based on a database of compound terms that are generated from the document corpus itself. There is no single magical algorithm for this, but a number of techniques combined that are now known as conceptSearching.

Concept based fuzzy phrase matching search

Whilst the search process is based on keywords, the ranking process is based on an index of compound terms. The phrase "capital gains tax" would be indexed as it is read, so the ranking process would rate documents containing the phrase higher than documents containing the individual words, even if the individual words appear with greater frequency than the phrase. The Fuzziness comes from the use by the ranking algorithm of word stems. Therefore, documents containing the phrase, but using "gain" or "taxes" would be treated the same in the ranking process.

A conventional approach to achieving greater precision in a search is by the use of exact phrase matching, usually denoted by adding quotation marks around the search term. Whilst this does indeed give a more precise (high precision) search, it will ignore the stemmed words and documents containing a high frequency of the individual words. Other high precision techniques involve searching only specific metadata, such as the document titles, authors etc or specifying word or paragraph proximity. These techniques belong in the realms of search science and simply will not be tolerated by today's information surfers.

conceptSearching's compound term indexes and fuzzy phrase matching attempt to provide high precision with high recall. The indexing engine automatically identifies multi-word concepts, and the results ranking algorithm assigns an appropriate weight to these compound terms. Documents containing the required concepts tend to be listed above those that simply contain the required words.

To understand how this might work, we will use the following example:

Consider the phrase "dangerous dog attacks baby"

The following two documents are both 'considered relevant' without concept based searching as they contain all the four key words of your search

dangerous dog caught attacking baby
dangerous virus attacks baby dog

The first sentence contains two (fuzzy matched) phrases contained in the query, the second sentence only one. Using a *wdf* approach, both documents would be ranked equally. Using conceptSearching we can see that the first document would be ranked higher.

Contextual suggestions

After a search is performed, researchers often refine the results by submitting a second search based on the results of the first query. This requires some knowledge of the topic (perhaps gained by the initial search) and then an ability to refine the original search terms. concept Search engines generate statistical profiles of documents during the indexing process. This can be used to guide the researcher by analysing the top hits returned and retrieving the most popular unique phrases contained in the hit list. Using this on-the-fly process conceptSearching extracts the "most nearly unique" concepts from your current hit list and presents them as 'related topics'. Researchers can select related topics of interest and use them to refine their search.

This can be compared to the Google approach where two searches are performed for each user search. The first is against the website index using PageRank™ relevancy, the second against a list of advertisers who want to attract people interested in the subject searched, with the highest bidder being the most "relevant", the theory being that if you were looking for that, you might want to

https://doi.org/10.1017/S1472669607002071 Published online by Cambridge University Press

buy this. Whilst related topics are generated from the same index as the documents, the effect is similar: the terms in the user's search produce a list of documents. The related topics will produce a similar list of documents, but with a different emphasis. This list may be more accurate than the initial search, or it may highlight concepts within the index of which the user was unaware.

For example, a search for "tank" may provide: oxygen tank; Chieftain tank; and septic tank as 'related topics'. By refining the search with the phrase oxygen tank, the secondary search will contain many oxygen tank related hits and few about septic, or Chieftain, tanks. The principle is that the related topics are dynamically generated by the conceptSearch engine through statistical knowledge of the entire document contents of the unique hit list. The engine treats common concepts within the main document corpus as noise and shows only concepts that appear more frequently in the initial search than in the general document corpus. The overall effect is that each set of related topics is tuned to the individual search without the need to teach the engine what concepts users are searching for. Additionally industry standard vocabularies can be added as a post-search sanitisation process (controlled vocabulary) to highlight only acceptable terms, such as legal acts and citations, or just to get a neater list of related topics. From an end-user viewpoint, they are given the opportunity to select from a list of "known good" related topics that provide intelligent refining and are far more convenient than having to think up search terms that may or may not exist in the document set.

Federated search and conceptSearching

Within the KM world there are many subscription-based sources that will not allow search engines to access their data. This proprietary information will almost always come with a complex Boolean search application that is often under-utilised by those who would benefit most from the content. In this scenario we can use XML Web Service technology to interface directly with the third party search engines to give the appearance of a single search crossing multiple repositories.

One example where all the above is brought together is Sysero (from UCLogic). This is an application that can aggregate industry specific web sites, internal and external facing information sources plus external information services providers, such as LexisNexis, and provide all the capabilities as described above.

Real world applications

The Sysero information portal from UCLogic uses conceptSearching to index documents and data held on internal servers and to aggregate external public web based information. Information is grouped by subject area

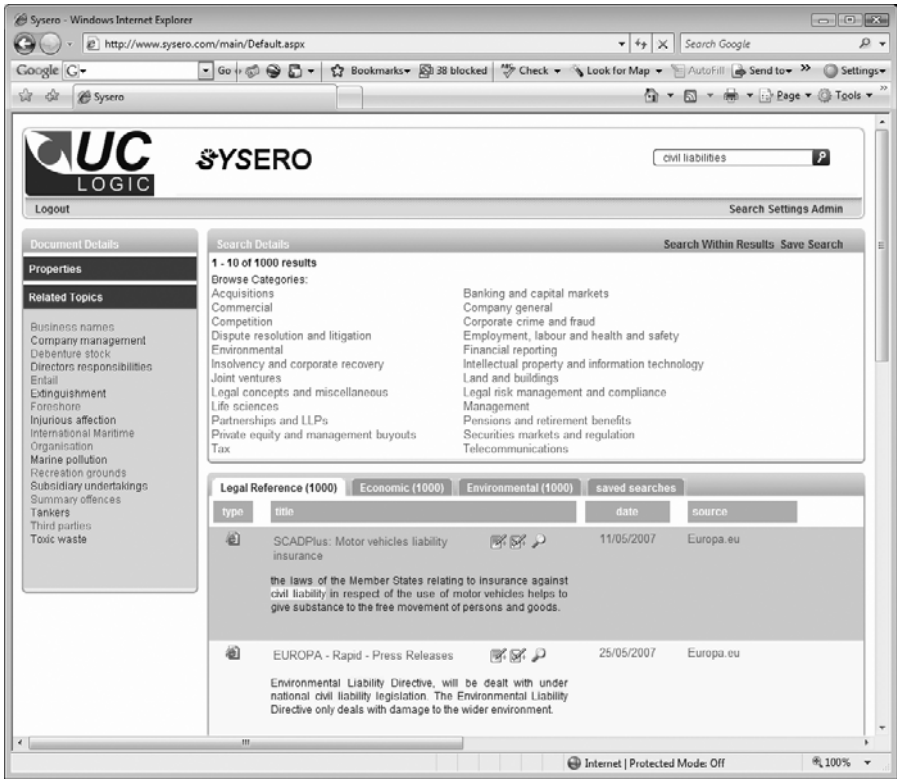

Figure 1: Logic Sysero Search details

https://doi.org/10.1017/S1472669607002071 Published online by Cambridge University Press

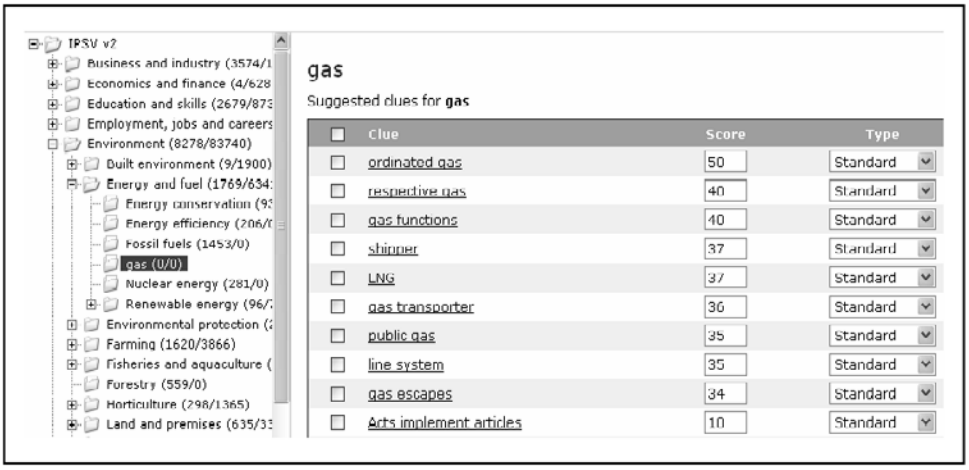

Figure 2: Taxonomy node management

(legal reference, economic, environmental, etc) using tabs. Additional tabs provide access to external 'paid for' information sources. Using Asynchronous JavaScript and XML (AJAX) technologies to overcome the differing speed of response of the various information sources, the various tabs highlight the number of results from each information pool at the point of time when the engine has returned the hit list. This single area of technical functionality is the main reason why applications like Sysero are succeeding where previous "global search" projects have failed. In a world where users expect search applications to deliver results within a few seconds, applications that need to wait for the slowest data source to deliver will appear slow and cumbersome, and will soon be rejected by the user community.

Other applications for conceptSearching

Modern search engines can also be used against semi-structured data to add value to research. One such example is Know Who searching. In Sysero, all document queries can also be run against an index taken from time and billing information. In professional service operations, time and billing systems capture time spent on work items. The core data in these systems describes the name of the individual carrying out the work, the time spent on the subject and a narrative describing the work product. By applying the document query against the index, conceptSearching can return the most relevant time entries and an XML processor can use grouping and summing to provide a list of those who have the most experience on the topics searched.

Formal classification systems

The ability to drill down through a taxonomy and view documents by subject area has a number of benefits over simple searching. Firstly, it uses a browsing metaphor where users are given suggestions of where to look before starting. Taxonomies are inherently context-sensitive, as there must be valid documents for the classification, or node in taxonomy parlance, to exist. This goes some way to achieving one of the goals we mentioned earlier of having the source documents guide the user through context. Secondly, the user has a smaller and more relevant list of documents to search through. In a poly-nodal taxonomy, researchers may find additional nodes attached to documents that may guide them to other areas of the taxonomy and further relevant material. The most noticeable effect of taxonomy, however, is to create a folder based storage analogy which mirrors the familiar directory structure approach used by PCs.

Formal taxonomies are not without their problems. Firstly, organisations have to build, buy or steal the basic outline of taxonomy. For a law firm this might look fairly straightforward, starting with the individual practice areas as the root nodes and fee earner expertise as the first child node. Further subdivision could be achieved by the fee earners applying their domain expertise to their first level nodes. In reality, though, this level of co-operation would require a huge organisational commitment and so we go back to pre-defined taxonomies, such as those sponsored by governmental bodies or commercial information providers.

Once the taxonomy outline is built (or more likely bought), building the rules that define a document's inclusion in a node is the next step and dwarfs the task of taxonomy creation. A number of organisations provide services where they will use legally skilled professionals to categorise documents, but this is clearly expensive and time consuming. This is an area where the statistical approach taken by conceptSearching techniques can help.

As discussed previously, the conceptSearching related topics facility extracts terms which define documents in an original, or seed, search query. Apply this to

https://doi.org/10.1017/S1472669607002071 Published online by Cambridge University Press

taxonomy rules management and we can use the taxonomy node itself as the seed query and pick the most useful related topics to define node inclusion. Further refinement can be achieved by filtering the related topics through an approved list of terms.

By providing an automatic suggestion process, conceptSearching taxonomy management reduces the need for taxonomy experts, as the automatic suggestion process uses the document corpus to find appropriate terms to populate each node on the taxonomy. To work within the technical parameters of the conceptSearching engine, the suggested terms are generally two and three word phrases, therefore the classification process of attaching documents to each node is more reliable. Large information providers and defence organisations who have used this method have noted an 80% improvement of taxonomy node maintenance.

Taxonomy management

To understand this in more detail, we have provided a Taxonomy Manager application. In the screenshot above we have created a new node called Gas. In order to create the rules for document inclusion, we have used the related topics ability to generate a list of unique terms from the documents which contain the word Gas. In order for a document to be included in this node it must score above an arbitrary threshold value, in this case set at 100. As the *wdf* of these terms is held within the index, a mathematically defined value can be calculated by the application for each set of related terms for the node. This in effect creates a list of synonym terms and a weighting. In conceptSearching terms these are called a clue and are part of the inclusion rules for a node. If a document contains enough clues to exceed the threshold, it is included in the node.

Whilst this approach is not totally automatic and requires a degree of subject matter expertise, it is a non-programmatic approach to developing a taxonomy. Additionally, the clues are generated from a statistical analysis of the documents indexed and therefore the rules directly reflect the content and are subsequently much better at classifying the documents into the correct node. Lastly, this approach uses the stemmed version of the words by default, although this can be overridden for individual clues, which means that documents do not need to have the exact phrases within them, just a close match.

Conclusions

Firms need to develop strategies to weed out the important data sources for their practice areas. One of the tools for this is an in-house search facility that is looking at the relevant data and is capable of unearthing relevant content without a scientific understanding of the underlying document sources. conceptSearching and applications such as Sysero are helping KM professionals build platforms that will keep their firms in the know.

Biography

Authors: Paul Billingham, Concept Searching Ltd (paulb@conceptsearching.com) and Phil Ayton, UC Logic Ltd (phil@uclogic.com).

https://doi.org/10.1017/S1472669607002071 Published online by Cambridge University Press