THE SCIENCE BEHIND PRF QUERY

Published by

Search Patents Real Fast™

www.patentsrealfast.com

PRF QUERY SCIENCE

Although the details of the PRF Query algorithm are protected for corporate competitive reasons, a brief overview of the basics of the process can be very helpful in understanding its value. Since the issue being addressed is how to search online for information in a better way, let's start with some general comments about how to achieve success using a search engine. For any search engine, the quality of results is influenced by the quality of the input query. Most often, while looking for new information, a user follows an iterative process of modifying his query depending on the current results from a previous query. In cases where the user is familiar with the available information, a good query is formulated resulting in good recall and precision of the search engine. In cases where the user's knowledge of the domain is limited and the number of stored documents is large, generating efficient queries may be quite difficult.

In addition, when the subject of the search is fairly complex, a query is usually continuously modified by the user to obtain good results. This is particularly evident when using a patent search site where the patents are complex documents, often describing multiple concepts in hard to understand language and using highly specific domain jargon. It may also be the case that those doing the search are not domain experts.

Finally, most of the large patent search database sites require the non-expert user to identify the important keywords and then use its unique command syntax to submit in a query. Their search tools are typically Boolean keyword based. Thus, the user must know the special syntax employed by these sites to create a valid query.

So, two of the main technology challenges for PRF Query were to help the non-expert to do the following:

1. Identify and rank the relevant keywords

2. Automate the building of the search scripts for the major free databases

To accomplish the first challenge, the PRF Query algorithm is built upon a statistical scheme that is focused on extracting the relevant keywords from a single document. The best source of information about a document is contained within the document itself.

The main idea of the PRF Query keyword algorithm is that, if a term occurs frequently with a set of other high frequency terms within a specific document, then this term is likely to also be significant. This co-occurrence bias is calculated using statistical tests. The tests are used to calculate the deviation of expected frequencies from observed frequencies.

The PRF Query approach is not the same as the tf-idf measure. The tf-idf measure promotes words that occur frequently within a document but occur less frequently in a document collection. This is often called the "bag of words" approach. This tf-idf metric is widely used in indexing schemes for search engines but many major search engines such as Google use other indexing schemes. If one can have the knowledge of the index of a search engine, then it would be easy to select the keywords for the query. But it is not feasible as the indexes change continuously over time as new documents are added to the collection indexed.

Thus, PRF Query utilizes the keyword co-occurrence algorithm which does not require any knowledge of the entire document collection or the indexing scheme used by any particular search site. With this algorithm, PRF Query is able to generate a basic set of important keywords from a single given text document submitted by a user. Once the set of initial keywords is generated from the document, the user may add words and information in two ways. First, additional keywords of the user's own choosing may be added to those automatically extracted from the text document. Second, the user may select date ranges and data search fields available at the selected specific search sites.

After extracting the important keywords, PRF Query is ready to accomplish the second major challenge: to automatically create the correct query script. Each of the major free databases uses its own syntax for their search query scripts. The syntax is not impossible to learn and many experienced patent searchers have become experts in writing the correct syntax for the specific database. But for the non-professional patent researcher, this can be a cumbersome and time consuming issue. PRF Query, though, makes this quick and easy.

From the keywords and the other selected query values, PRF Query generates a syntactically correct query string to submit to the selected search site. And, as the user engages in the iterative process of selecting different keywords or other search options, PRF Query automatically adjusts the search script accordingly.

