From Keyword Search to Conceptual Search

March, 2006

No matter how celebrated search engines such as Google and Yahoo are, and how indispensable they appear to be, search technologies are very limited when it comes to searching vague, but not necessarily complex, *concepts*. Today's search engines work fine if the *concept* you want to find can be represented or described by unique keywords. Here are some examples:

You want information about Sony Walkman, or Grand Hyatt, or … These are unique names that have precise meanings. Type them into the search box, and there is a very good chance that you will find all of the results to be relevant or even useful.

Another example: you want to find George Bush. Search engines will quickly give you very good results about George Bush, whether he is the President of United States or one of your high school buddies, you will find it there (of course, you need to scroll down dozens and dozens of pages before you find your buddy's page). In this case, "George Bush" refers to only a small set of objects in this universe.

Any thing slightly more complicated is not so easy. For example, you want to find information about software that can screen out data that match certain patterns. You can see that this "concept" involves *at least* three sub-concepts, and each can be described by many possible keywords:

The object:	data, information, record, …
The action:	screening, filtering, matching, …
The mechanism:	software, system, server, …

For example, what you want to find may be written as "data screening server" on one page or "information filtering software" on another. The table above immediately yields 27 combinations, not including spelling variations such as "systems" and "servers" which are treated differently by most search engines. Examining a few result pages for one query is time consuming enough; now, imagine doing it 27 times or more.

Of course, if you just want to find one or two results, say, to use in the bibliography of an article, you may get enough by trying just one query. However, if your goal is to research the topic thoroughly, you need to try many "equivalent" queries – and the worst thing is that you have no idea if you have missed any important combinations.

Although all of the 27 combinations are plausible for describing what you want, their usages in web pages are quite different. Searching these 27 queries on Google yields the following (performed on Feb 19, 2006):

data screening software	88
data screening system	136
data screening server	0
data filtering software	289
data filtering system	188
data filtering server	2
data matching software	269
data matching system	340
data matching server	1
information screening software	2
information screening system	9
information screening server	3
information filtering software	363
information filtering system	11400

information filtering server	5
information matching software	7
information matching system	395
information matching server	0
record screening software	12
record screening system	2
record screening server	0
record filtering software	6
record filtering system	4
record filtering server	4
record matching software	162
record matching system	154
record matching server	0

There are a few observations/surprises:

- The term “server” is extremely unpopular in this domain although most application software runs as servers, as in web server, application server, and database servers, etc., the nine queries using the term “server” return a total of only 15 results (four of the nine queries return 0 results).
- The term “system” in the queries consistently returns more results than the term “software” in the queries (with only one insignificant exception), although “systems” are in fact “software systems” and as such “software” is a more precise term to use.
- Although the term “data” is much more commonly used than “record,” “record matching software” and “record matching system” return respectable result sizes of 162 and 154, respectively.

The conclusion of this article is that today’s search engines require you to supply the suitable queries. Since there is no way for you to come up with and try out all possible expressions of the same concept, you can only expect to find a few results but not the best and comprehensive results – imagine what you would have missed if you had not tried “information filtering system.”

Search engines of the future should do it for you, automatically and transparently. By expanding the query semantically and ranking the results based on how close they are to the original query, a search engine can return results that are both broad and yet precise. If you have any question about this article or want to discuss the solutions to this problem, please email suntek@suntek.com.hk.