Linda Pagli

Search Engines

Abstract

The so called *Search Engines* are basic tools to search and retrieve information in the World Wide Web. They are complex systems of hardware and software components able to return, after the specification of one or more *keywords*, the web pages, found in their archives, containing that keywords. Special programs, called *crawlers* o *spiders*, always active, are sent by the search engines around the network; they travel from page to page following the pointers, select the relevant information, and bring it back to the engine. The collected information is compressed, stored and indexed in the memories of the search engines; the incoming requests are then processed into these enormous indices of keywords.

The first search engines did not function very well: after a query one was submerged of useless or meaningless information, from which it was very difficult to extract the searched information. Only expert users, capable to express the requests in formal way could be successful, for all the others the sea of information was not navigable. The problem was to find a technique able to select, among all the pages containing the keywords, the relevant ones in an automatic (without the human intervention) and cheap way. Google gave the first solution to this problem. Two young ph-d students of Stanford University, Sergey Brin and Larry Page had the right idea to evaluate the popularity of a web-page through the number of other web pages pointing to it. Their idea intuitively was: if a page is very popular it is probable that it is interesting for many users, and in particular for the one who submitted the query. They defined mathematically a score, called *page rank* to be assigned to each web page and easily computed.

The page rank of a page P is computed as the sum of all the page ranks of the web pages Q's pointing to P, each one divided by the number of its outgoing links. This means that the more a page is popular in the web the more will be its influence on the web pages it refers. Its influence also depends inversely on the number of its outgoing links. A page rank is assigned to the pages containing the requested keywords, then the pages are sorted in decreasing order of this rank, and this list is produced by Google as result of a query. The page rank is not the only parameter considered in order to produce the result, there are many more, but they are kept secret and probably updated from time to time. With this method it is not certain that the highest score page is the one with the best quality answer, but in most cases it works quite well, and after Google, people have started to use the information in the web in a more satisfactory and effective way. Other ingredients of Google success are: the very basic interface with no banners, the huge set of servers and the additional new services, such as Google Scholar directed to researchers of the academy, and many others. After Google, also the other search engines improved remarkably their service and the life of the web *surfers* became easier.

One can reasonably ask whether the lists produced by the search engines can be manipulated and the visibility of the web pages artificially increased. This is currently done for many reasons, but mainly for commercial interests. The technique is called Google-bombing, it is well known and consists in automatically producing a huge set of web pages containing only one reference to the target page and the keywords that have to be associated to this page. There are also other methods that include frequently searched keywords or sentences, hidden in the text of the web page. Companies have started to sell visibility on the web, thus provoking the so called *bubble* of web visibility. Suppose a set of pages matches a given query, if all page owners want to appear high in the resulting list the competition becomes an unstable process in which page owners buy additional visibility. This bubble reminds us of a phenomenon typical of the stock market. There is a continuous battle between search engine operators and companies for web pages promotion, in which search engines periodically modify their ranking model; this battle can be easily observed following from time to time the list associated to the same set of keywords. However the web belongs to everybody and these manipulating efforts become the target of sustained protest.

The new generation of search engines try to overcome some of other limits, such as those related to keywords with several meanings: in this case the answers mix together many undesired web pages and the good ones,. One way to address this problem is to design search engines capable of retrieve documents “relevant” for the user, by taking into account some contextual or profiled information (Eurekster, Yahoo, Google, AskJeeves). Other new search engines, called meta search engines, work on top of normal search engines. The results are clustered into folders, labeled with sentences sharply derived from the answers. The folders are then organized in a hierarchy (Clusty, SnakeT), thus overcoming the above problem. Other problems related to the quality of the answer, reflect more the quality of the information stored in the web, than the limits of the search engines.