Categorized and Controlled Search Engine – A Prototype

Shmuel Gutman

Abstract:

During the development of the Internet over the recent years, searching for information became one of the major applications for Internet users. Although improvements made in the search engines regarding the scope and ranking algorithms, users still encounter several major problems, reflected by search results which are not fully compatible with their needs. Sometimes, the search results are not related at all to the term the user searched for. Sometimes the search results do indeed contain the search words the user asked for, but in a totally different context. Additionally, usability studies conducted points out that the average user checks only the first few results of the search engine, so when relevant information is not on the top of the search results list, the user is not exposed to it at all. These problems often cause the user to give up on finding the information, or in worse cases, leaving the user with partial or mistaken information. On many other occasions, the user is forced to try several different queries, repeating the entire process several times until finding the required information, consuming a great deal of time and effort.

The goal of this research is to give a solution to these problems and create an innovative search system that will give the user the ability to quickly focus on the quality search results in the field relevant to him. In order to achieve these goals, the system will utilize existing search engine and web directory by integrating data from both of them, in a way that search results received from the search engine, will be displayed to the user arranged by group of topics based on the their cataloged information as it stored in the web directory. Using the system, the user can be quickly targeted, with much ease, at the requested information, laying out in front of him the quality and controlled data in a clear way.

This system will be called Category/Controlled Search Engine or CSE in short. In its core is an innovative algorithm. This algorithm uses the Google

search engine in order to retrieve information from the internet and Yahoo web directory in order to receive information regarding the result retrieved. The algorithm quires the search engine with the requested search term and then consecutively pass over the search results web addresses (URL) retrieved from Google search engine, checking under what category it is located in Yahoo web directory. If the exact URL is not found in the web directory, the algorithm tries to locate a partial URL or at least the topic of the site in which he document was found. This way, the algorithm provides, for each search term, a collection of results from the search engine, where each result has additional information regarding the category to which this result belongs to. The output of the running algorithm will be the display of the gathered search results and cataloged information, in a way that slightly resembles the existing clustering engines, where each result is located in a categories tree based on the information received from the web directory. This way, the CSE harnesses, from one aspect, the text search scope of search engines, and from another aspect, the human understanding and inspection and the division of topics found in the web directory. Likewise, this system will take advantage of the clustering method for displaying information, particularly the improved retrieving ability of the user by focusing on the desired topic, without the need to go over dozens of irrelevant results and topical distribution of the desired field of interest. The advantage of the CSE over classic clustered engines is that the distribution topics are based on human supervision rather than on statistical textual analysis.

The research is composed of three stages:

1. Locating tools and resources required for conducting the research.

2. Development and initial trial of the system for probability checks and system enhancement.

3. Evaluation and trial of the system against exiting search engines.

The CSE algorithm utilizes an existing search engine and web directory. Therefore, at the first stage the need to locate an adequate search engine and web directory, suiting the research needs arose. Since the research aims for

the average user, the search engine found adequate is Google, considered to be the most popular search engine in the world. The web directory found adequate for the research is Yahoo which excels in size, scope and clear hierarchical structure. In order to evaluate the CSE, arose the need to locate an existing clusters engine. The clusters engine finally chosen is the carrot2.org, mainly since it is the only clusters engine that exposes the algorithm at its core and operates under free license. For the trial itself, the 50 most popular queries, reviewed in statistical publications, were used.

For development and trial of the CSE, an application was designed and written in order to implement the CSE algorithm. The application tested for 10 search terms, for each 100 results were retrieved from Google, bringing it to total of 1000 search results. For each Google result, information gathered from Yahoo about the result's category. Analysis of the results indicates that even though there is a major gap in the order of magnitude between Google results count and Yahoo directory results count, Yahoo provided information on the majority of Google's results.

In a trial conducted, about a quarter (25%) out of Google's search results were found on Yahoo directory with exact match. When only the top 100 results, considered to be with highest quality, were checked, the number increased to 40%. When partial matches counted, comparing the domain names of the search results, the number increased to 70%. Analysis of the URL of the results indicates that more than third of the results retrieved by Google are domain names and 7% of the results retrieved by Yahoo directory are internal page in a site. This finding weakens the claim that since a search engine retrieves pages, on contrary to web directory that classifies sites, there is no way to study search engines relying on web directory. A manual test performed on 20 results out of the 1000, with the purpose of finding out whether the information retrieved from Yahoo assists in correctly classifying the search engine search results. It turned out that in 70% of the cases, the category found in the web directory, does contribute to correctly classifying the search engine results. In 55% of the cases, not only the category correctly classifies the results, but actually assists in retrieving the information, by

adding additional information on the result found. These findings indicate that the information from web directory can assist in classifying the search results. In an additional trial conducted in the research, the categorized information retrieved from Yahoo was checked. The research showed that in over 70% of the categories received, there is no duplication, which in turn, made it difficult to categorically gather the results. As a result, it was decided that the categorically gathering of the information will be done based on the hierarchical position of the category rather on its name. This way, a sample category result in the category tree located under News_and_Media/Newspapers will be gathered with a result located in News_and_Media/Magazines.

In order to take full advantage of the CSE and achieve the objectives of the research, an interface, somewhat resembling the clusters engine interface, was developed, by utilizing the Yahoo directory hierarchical structure and adjusting it to suit our goals. This way, the results were concentrated under fixed categories of topics. Each category selected by the user, revealed him with results matching that category along side matching subcategories. This way the user is quickly exposed only to the supervised results in the topic of his interest, without the need to select from the entire results list.

Further more, the user is exposed to results that he probably wouldn't have reached in the absence of the system, since the classic ranking oriented search engine would have listed them relatively low. The Yahoo hierarchical structure was not entirely duplicated, but partially, due to its massive scope and deepness.

The system was developed in C# on Microsoft.NET 2005 platform and it is currently running on a Windows 2003 server at http://67.192.98/cse.

In order to measure, compare and evaluate CSE, several indexes were defined for the time and quantity of results the user had to go over until he found the desired information. Later, a trial was made in order to compare the indexes marks resulted from the three search engines (Google, Carrot2 and CSE) for 9 terms in different topics. The trial results indicate that for the average user, the CSE system provides better results than Google. The

system also provided the information, by average, five times faster and the user had to go over on only fifth of the results, comparing to Google. Only when the user wants and able to build a detailed search term, composed of three words, and has prior knowledge on the topic, the Google search is preferable. The CSE system also showed better results compare to the carrot2 clusters engine when the user lacked information on the search term.

This research contributes both in research and implementation applications in both search engine and search engines evaluation fields. The research suggests a prototyped and innovative search engine that combines the benefits of a search engine and web directory, thus ensuring the user quick and quality results.

Finally, the research suggests improvements to the CSE in order to overcome the shortcomings discovered, thus turning it into a leading and innovative search engine in its area.

System no. 1153582
025.04 גוט.מנ תשס"ח