Discussion Report: Hypersearching the web

Azbayar Demberel
Department of Computer Science
asic@cs.duke.edu

January 27, 2007

1 The challenges of search engines

With the WWW growing hour by hour, searching and obtaining high quality relevant results from the collection of unstructured/unorganized information (which WWW trully is) has become a complicated task. In addition, web sites are written in multiple languages, styles and dialects, containing truth, falsehood, wisdom, propaganda or sheer nonsense. Distinguishing the most relevant information from the thousands of others who contain the exact same keywords but completely different context and aim is both challenging and important.

Search engines use heuristics - also known as ranking functions, to prioritize and hence determine the relevance of web sites in regard to the search term. In the past, search engines implemented simple heuristics like favoring pages by the number of times they contain the query term or by the location and size of the keywords. Simple heuristics, however, did more harm than help to the search results. As these heuristics are easy to be manipulated, many commercial web sites used to exploit their weaknesses using techniques like spamming, which made it very difficult to maintain an effective search engine. For example, they could insert phrases many times over in colors and fonts that are invisible to human eyes. The search engines with simple heuristics, however, will count all of the words as valid and would give the web page a favorable ranking.

Moreover, human language, rich with synonymy and polysemy, makes the search even more complex. For example consider a word like “business”. It can have multiple meanings[1] like (a) a purposeful activity, (b) a role or function, (c) an affair or matter, (d) a personal concern, etc, and it can be expressed or substituted by many different words like commerce, trade, industry, work, etc. So whenever a search is performed, it not sufficient to return the results of only the keyword. In fact many of the more relevant web pages might not even contain the search keywords. For example for a search for “automobiles”, many pages might lack the word “automobile” but instead contain “car” and such results of course cannot be excluded.

[1] Merriam Webster dictionary.http://www.m-w.com/

To solve such problems, administrators tend to intervene to the searches of their search engines. They do so by hard coding some results which they believe they know what the appropriate responses should be like. In other words, they predetermine the “right” answers and over-write whatever results their search engines produce. This approach, although workable for a certain extent, is totally unscalable and highly subjective. There are countless possible queries and maintaining a list of precomputed results for each of them is simply impossible with the current growth of the web.

2 Solution to the problem: Hyperlinks

Hyperlink is one of the core elements of the WWW and is the main tool to navigate between web pages. Since people use the hyperlinks to reach to their desired web sites and the goal of searches is to return users their “desired” results, it is evident that hyperlinks should have a major role in ranking web pages.

Using hyperlinks to order web pages in terms of their relavance to the search query is an idea similar to the scientific paper citations: A scientific paper's importance is determined by the number of citations it receives. This link analysis however, cannot be directly implemented in the web. For example, if the impact factor would correspond to the ranking of a page simply by the number of links that point to it, it can favor universally popular locations, such as the home page of the New York Times, regardless of the specific query topic. In addition, a link in a web page, in contrast to citations in scientific papers, may also exist for navigational purposes only, e.g. “Click here to return to main menu”.

If a smart algorithm using hyperlinks can be created, with a mind on their shortcomings, it has a potential to become a very efficient heuristic.

3 Clever's approach of clever searching

Clever uses hyperlinks to distinguish 2 types of web pages: authorities - the core important sites for various different topics, and hubs - web pages that point to those important sites. In a sense, a respected authority is a page that is referred to by many good hubs; a useful hub is a location that points to many valuable authorities. For every searches, Clever finds the respected authorities by using the corresponding hubs on that topic using the following algorithm:

1. Get a set of candidate pages about the topic.

2. For each one make a guess about how good a hub or authority it is

3. Use the current guesses about the authorities to improve the estimates of hubs: locate all the best authorities, see which pages point to them and call those locations good hubs

4. Take the updated hub information and refine the guesses about the authorities: determine where the best hubs point most heavily and call these the good authorities.

5. Repeat steps 3,4 while necessary

Clever obtains a set of candidate pages from a standard text index such as AltaVista, which is usually a list of 200 pages. The system then augments them by adding all pages that link to and from that 200. The resulting collection, called the root set, was typically between 1,000 and 5,000 pages. In steps 3 and 4, a page that has many high-scoring hubs pointing to it earns a higher authority score; a location that points to many high-scoring authorities garners a higher hub score. Clever repeats these calculations until the scores settle on their final values, from which the best authorities, i.e. search results, and hubs can be determined.

The authors proved the validity of their algorithm using algebraic analysis. They have also shown that the iterative process (step 5 of the algorithm) rapidly settled to a steady set of hub and authority scores. For example, for a root set of 3,000 pages, it required five rounds of calculations.

An interesting and useful aspect of Clever was that it naturally separated web sites into clusters of similar subjects. A search for information on abortion, for example, resulted in creating two groups of web pages: pro-life and pro-choice.

4 Future work

The authors noted 2 major objectives in their paper.

Integrate text and hyperlinks One way to improve search results is by weighing links differently based on the relevance of the text of the hyperlink and the surrounding text. If the query text appeared frequently and close to a link, for instance, the corresponding weight would be increased.

Construct lists of web resources By constructing lists of web resources, similar to those of Yahoo! and Infoseek, the authors think they can uncover hidden small communities of web pages that don't interract with other web sites. They also argue that such automatically compiled lists can be competitive with manually created ones.