AN OVERVIEW OF SEARCHING AND DISCOVERING WEB BASED INFORMATION RESOURCES

Cezar VASILESCU

Regional Department of Defense Resources Management Studies

Abstract: *The Internet becomes for most of us a daily used instrument, for professional or personal reasons. We even do not remember the times when a computer and a broadband connection were luxury items. More and more people are relying on the complicated web network to find the needed information*

This paper presents an overview of Internet search related issues, upon search engines and describes the parties and the basic mechanism that is embedded in a search for web based information resources. Also presents ways to increase the efficiency of web searches, through a better understanding of what search engines ignore at websites content.

Keywords: *information resources, search engines, spiders, web pages, information system.*

1. INTRODUCTION

The Internet becomes for most of us a daily used instrument, for professional or personal reasons. We even do not remember the times when a computer and a broadband connection were luxury items. More and more people are relying on the complicated web network to find the needed information. But how do we find that information? Like as we search a local database using specialized software tools, the Internet search is done in almost the same manner. Let's consider the Internet without search, with no means to navigate. As a result, the web as we know does not exist. We introduced so far two notions: the Internet and Search engines.

We can define the Internet as "*a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide. It is a network of networks that consists of millions of private, public, academic, business, and government networks of local to global scope that are linked by a broad array of electronic and optical networking technologies*" [1].

Another definition that is close to our goals states that *"the Internet is an electronic communications network that connects computer networks and organizational computer facilities around the world"* [2]. Finally, a third definition states: *"a global computer network providing a variety of information and communication facilities, consisting of interconnected*

networks using standardized communication protocols" [3].

A search engine is "a program that *searches for and identifies items in a database that correspond to keywords or characters specified by the user, used especially for finding particular sites on the Internet"* [3]. Another definition said that it is *"a program that searches documents for specified keywords and returns a list of the documents where the keywords were found"* [4].

Even if in reality search engines is a general class of programs, the term is often used to specifically describe systems like Google, AltaVista or Yahoo search that enable users to search for documents on the World Wide Web.

Why is searching the Web necessary? The answer is that the Internet is too large and chaotic to find much useful information. The initial web architecture was designed for the purpose of simply interconnect pieces of information. The term hyperlink means that the searcher could only find information manually, moving from one piece of information to another by following hyperlinks connecting one page to another.

Search engines aggregate, concentrate and organize information, following a huge number of hyperlinks and collecting information for later retrieval. Since no statistics are recently available, we have to rely on Google's 2005 announcement that their search engine surpassed 8 billion indexed items (*"Searching 8.168.684.336 Web pages"* [5]) or on Yahoo's response that "*... it had the largest search-engine index, tracking 19,2 billion documents"* [5], but the only reasonable way to make an accurate and independent count of the index sizes is to find an objective noninvolved third party.

The search for web based information resources involves three parties:

a) the searcher (person and it's computer)
b) one or many search engines, and
c) the web site searched.

It also requires extensive use of the Internet infrastructure and protocols in order that the three parties to communicate. It would be interesting to present a simplified network configuration necessary for connecting to and searching the Internet (figure 1).

The searcher's computer connects to the Internet through an Internet service provider (ISP) that usually offers the address resolution service using a domain name system (DNS) server. The DNS server is necessary to translate site names (such as www.dresmara.ro) typed by the users in browser into the corresponding IP addresses used by Internet routers.

In order to receive the search engine's home page, a searcher must complete the following steps:

1. The searcher enters www.search.com in browser's address field.
2. The browser sends the request to DNS server (154.69.87.23) to translate www.search.com into an IP address.
3. The DNS server responds to browser with 167.89.12.34 address.
4. The browser requests that 167.89.12.34 send its default page (usually index.htm).
5. www.search.com responds to browser with the search engine's default page.
6. Browser displays search engine's default page; searcher can now enter query, sending a search request that would repeat the process again.

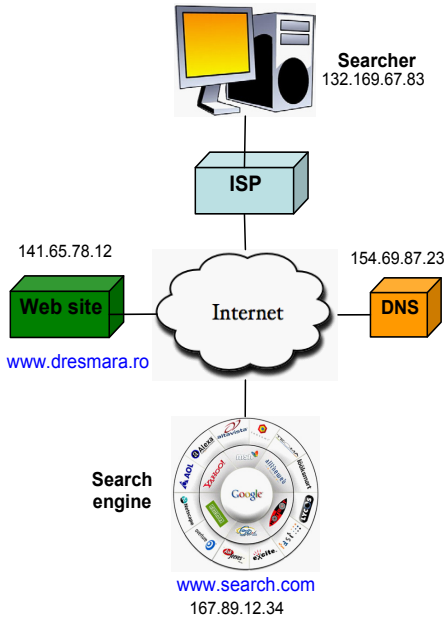

Figure 1. A simplified network configuration.

The mechanism of the process repeats identically for receiving the www.dresmara.ro home page.

2. SEARCH ENGINES

The primary purpose of the search engine is to close the information gap between the searcher and Web site pages. From a searcher point of view, a search engine purpose is to match the query words with words on Web pages and to list the pages containing the matching words in a relevant order.

In the typical search engine architecture, one search engine component called a spider (or crawler, robot, etc.) visits the Web site to retrieve pages linked from the main page just like a person using a browser would follow links to new pages. The crawler follows links, indexing meaningful words of the retrieved pages to be stored along with the page's Web site location in a specialized database on the search engine for later searches and retrievals [6].

As the searcher sends queries to the search engine, the engine consults the database for pages containing similar words and phrases. After ranking the pages for relevancy to the query, the engine returns the ranked list of Web site page locations. The searcher selects location entries from the list and retrieves pages directly from the Web site.

The accuracy and speed of page ranking calculating process improve the relevance and quality of the web based information resources.

A key point of competition between search engines is the way to rank one document against thousands of others. In reality, an identical search on multiple search engines will produce different results. Based on the search engine's „brand" the searchers chose one search engine.

In order to obtain best results, it is recommendable to consult several search engines to have confidence in the results. Meta-search engines aggregate the results and automate the process searching on multiple sites, by sending the user search query to several search engines and creating a fusion of the results.

3. WHAT SEARCH ENGINES SEARCH AND WHAT THEY IGNORE

In our days web design purpose is not only to create a good looking website, but also to create an "appealing" one for the visitors. Websites compete against each other for a high page rank that means a better position in the page that result after a query.

We often hear rumors about search

engines selling placement at the top, or companies buying high-ranked websites, because high placement in the ranking means money to the commercial Web site. That leads to several reasons to question regarding the accuracy and veracity of the ranking results.

Generally, search engines tend to improve themselves by extracting added information from web pages elements, such as:

- the title
- description
- key word HTML tags
- the meta keywords tag
- the text content section and
- the connecting links to and from the pages.

Most indexing spiders examine only the first few hundred words of content so it is important to early provide descriptive key words in the content text.

Knowing the parts of a web page that attract the attention of indexing crawlers is critical to website designers in their attempt to raise the visibility of their site. Because guidelines to website promotion were extensively published and analyzed, a website designer could proceed in a semi legal manner to "fool" the visiting spiders to index the pages and finally high rank their website.

Examples of manipulation could include web pages that contain:

1. Deceptive text
2. Intentionally misleading links
3. Deceptive self linking referencing patterns among others in a group
4. Off-topic or excessive keywords
5. Duplicate content
6. Different content than the spidered pages
7. Material designed to lead users to other web pages
8. Metadata that do not accurately describe the content of the web page.

A study of search success [7] illustrates the difficulties and necessity of designing a Web site designed for search. After watching 30 searchers search different sites for content that was on the sites, the study concluded, "we observed that users only found their target content 34% of the time with one search".

In order to increase the efficiency of our searches, we must understand what search engines ignore at a website.

Most spiders purposely ignore or cannot see large parts of a page that human readers might see and use, effectively hiding that part from search engines.

Every page designer expects the spider to index the full page and follow all links to other pages, but this is not happening all the time. By ignorance or bad design, spiders can be excluded from indexing some or all pages of a web site.

Here are some examples of what the search engines might ignore:

a) Frames

Can stop an indexing spider, because frames require at least three separate pages:

- a hidden frameset page that defines the frames and links visible pages into the frames,
- a second page for visible content, and
- a third page for navigation.

A spider usually arrives at the hidden frameset page, but in order to follow links to the other visible pages it must understand how to handle the frames.

Spiders that do not understand frames simply stop and never index further. For those spiders and browsers that do not understand frames, the remaining site pages are unreachable unless alternative links

exist.

b) HTTPS protocol

Spiders generally use only HTTP for communications with web servers and do not normally index a server requiring HTTPS.

c) Scripts

Most spiders ignore script programs written in JavaScript or other scripting languages, others simply index the script program text.

The reasons for spiders ignore scripts are: 1. they must be able to execute the script, which requires an interpreter for the script language; 2. spiders must simulate any required human interactions such as button clicks.

As a result, scripts can hide sensitive information, which sometimes can be good.

d) Java applets and plug-ins

Any browser-executed program is invisible for a spider since no text is provided, other than that needed to execute the program. Unless the spider is able to execute the program, there is nothing to index, and often if executed by the spider, the program output is most likely graphical and unreadable by the spider.

e) Server-generated pages

Some searches may ignore unusual link references that do not end in "HTM" or "HTML." For example, a spider will follow the link "<a href = 'mainpage.HTML'>" but may not follow the link to the web server program of "<a href = 'mainpage. ASP'>". If generating the main website page in this manner could result that some spiders completely ignore the site.

f) Forms

Some sites offer to visitors the possibility to fill out forms, in order to collect relevant information, such as feedback on different functions, impressions or comments regarding the content, etc. To complete a form the interaction must be among a human person and the form itself.

Spiders don't know how to fill out forms, and the result for the indexing and search function is the appearance of potential problems generated by leading visitors from a search engine directly to a form page.

g) Spider exclusion

To exclude spiders from indexing certain pages site administrators could use the unique (for the entire site) "robots.txt" file that lists acceptable and unacceptable directories, pages, and spiders.

Another solution is the usage of special meta-tags to specify how the spiders should index each individual page: index the page or not, or the page links should be followed or not.

Depending of the spider's purpose, these "rules" could be (or not) followed. For example, an e-mail address harvester will certainly ignore the exclusion instructions.

h) Images

Spiders may index the image location, image title, and alternate text but that is probably all due to the effort required to analyze an image.

i) Meta-tags

Most spiders ignore Meta tags due to several attempts to manipulate the sites rank. A common past approach was to include in the Meta tags a repeating keyword, over and over. As a result the site was artificially high ranked. A study [8] made in 2007 revealed that at that date Google didn't check the Meta keywords tags, but Yahoo still did.

j) Deeply linked pages

A common belief said that every page of a website is indexed by search engines. In reality, only small sites are checked completely, while larger ones had only a limited number of pages indexed. In practice,

it's recommendable not to have "deep links" because there is a link depth threshold that spiders do not cross, ignoring pages linked beyond that depth.

4. CONCLUSIONS

So far I described a few issues that must be taken into consideration when trying to search web based information resources.

Search engines could be a significant tool in information resources retrieval, but only when used by knowledgeable users. In this respect, here are some web search related problems that must be overcome:

1. Lack of knowledge on how search engines work;
2. Finding the exact topic: Getting too many hits and narrowing down the search;
3. URLs changing or disappearing or hard to memorize addresses;
4. Lack of awareness of browser features.

In conclusion, search engines are the most used search spots in the Internet because it categorize a large amount of stored data and could provide results even for unusual search requests. Knowing some insides of them could improve the quality of our searches, the resulting information and not ultimately our informed decisions.

REFERENCES

[1] *** - http://en.wikipedia.org/wiki/Internet
[2] *** - http://www.merriam-webster.com/dictionary/internet
[3] *** - Compact Oxford English Dictionary
[4] *** - http://www.webopedia.com
[5] *** - http://www.nytimes.com/2005/09/27/technology/27search.html
[6] Arasu A., Cho J., Garcia-Molina H., Paepcke A., Raghavan S. - *Searching the web*, ACM Transactions on Internet Technology, 2001
[7] Spool, J.M. - *Users Don't Learn to Search Better*, 2001, retrieved from http://www.uie.com/articles/learn_to_search/
[8] Sullivan, D. - Meta Keywords Tag 101: How to "Legally" Hide Words on Your Pages for Search Engines, 2007, retrieved from http://searchengineland.com/meta-keywords-tag-101-how-to-legally-hide-words-on-your-pages-for-search-engines-12099