Sinteza
Međunarodna naučna konferencija Univerziteta Singidunum

Impact of Internet on Business Activities in Serbia and Worldwide

Uticaj Interneta na poslovanje u Srbiji i svetu

DOI: 10.15308/SINTEZA-2014-47-51

RESEARCH METHODS AND INTERNET DATA: THREATS & OPPORTUNITIES

Miroslav Trifunović, Kirill Boldyrev
Singidunum University, Belgrade, Serbia

Abstract:
In the last decade, Internet has become the most important source of information for any type of business or scientific research. Historical troubles of finding particular information have ceased to exist. The new challenge is to find the right information among a vast number of sources available on the Internet. Internet search engines are the first hand help, however, the ways they work are closely guarded by the software vendors. The paper analyses the ways Google search engine is searching and ranking requested information.

Key words:
data integrity,
reliability,
search engines,
Google PageRank.

INTRODUCTION

In June 2009, Reuters news agency has conducted a survey that has confirmed that Internet is by far the most popular choice of information. The same survey has confirmed Internet as the most reliable source for news (nearly 40% of adults), ahead of TV (17%, newspapers (16% and radio (13%). With increased proliferation of IT technologies, broad-band networks, computer ownership and wireless devices, there is a little doubt that as an information source Internet is dominating other key information players such as TV, Radio and printed information media.

RESEARCH METHODS AND SECONDARY DATA

Every serious research project, either in the domain of social sciences or in the domain of natural sciences, has to start by study of the prior art. Collection and analysis of the secondary data (data obtained by other authors) plays a key role in many stages of the scientific research including finalization of the research question or hypothesis, formalization of the research proposal, development of the research design, budgeting and reporting. In some cases, the whole research task can be based on the secondary data, without direct collection of the primary information using techniques of experiments, surveys or observation.

For these reasons collection of the secondary data is critically important and has direct influence on the research results

Classical research methodology literature refers to indexes, bibliographies, dictionaries, encyclopaedias, handbooks and directories as main information sources for secondary data collection. Today most of these resources are available in the digital format on internet. In addition there is a vast quantity of information appearing on the individual expert web sites, blogs, and even social media. Such abundance of the information requires new skills to collect the relevant and reliable data and to include them in the research process.

The internet data collection poses two main challenges: finding the data and evaluating the data quality and reliability.

DATA FINDING & INTERNET SEARCH ENGINES

As a general rule, data finding on internet starts with use of one of the available web directories or search engines. While web directories are human managed listing of the web sites, search engines work automatically and without direct human intervention. They are a basic software tool with a capability to access and index a huge and ever-growing number of web-sites, proximately doubling every two years. According to Netcraft/Google statistics from 2012, in year 2000 Internet comprised 8 million web sites and in 2011almost 400 million web sites. Digital Strategy Consulting's research from 2013 has shown 634 million web sites and 3.7 billion internet users in 2012. Today, information on the internet is stored on estimated one billion of web sites and hundreds of billions of internet pages. Without sophisticated softer solutions, diligent use of such information would be impossible.

Search engines operation

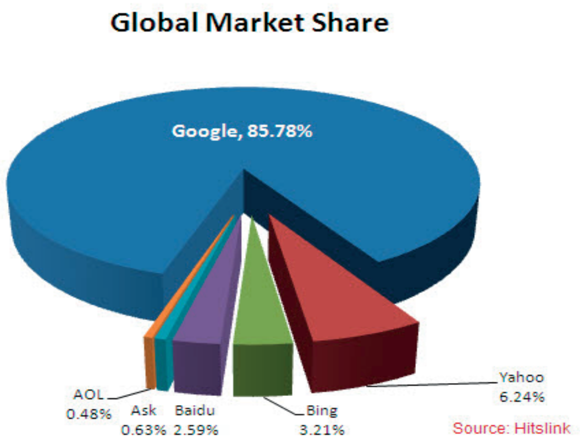
Computers with the internet connections are like houses with the open door. Through these doors information is entering from the internet to our computers or our information is placed in in the internet servers. Computer users are in control of such purposeful data exchange, however, they are not in control of the operation of the search engines that use the same paths.

Search engines are small software systems ("spiders", "crawlers", "web robots", "internet bots" or "bots") designed to visit web sites' pages, to search for particular information and to index the pages. They can be of a general purpose or covering only specific type of content / topic (e.g. patents, books, music etc.). They can be monolingual, bilingual or multilingual and have limited or unlimited geographical coverage. Their operation is automated and they typically work with one or more keywords that are the basis for information acquisition.

Development of search engines dates to early nineties, with rather small number of web sites and rather simple software tools to finding the data on Internet. During the last two decades the sheer number of the web sites and their increased complexity required the development of much more sophisticated and ever developing software solutions for data search and indexing.

Today's search engines

Today there are more than 30 major active search engines used to look for the specific information on the Internet. Clear leader is Google search with more than 1billion unique visitors / month. Different surveys (and measurement techniques) give to Google market share participation of 75 – 90%:



Picture. 1. Search EnginesMarket Shares
(source: http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4)

With such dominance and widespread use of Google in information search, it is important to understand how Google search engine works. Although the details of the software tool is a closely guided professional secret and the solution itself is being continuously developed and upgraded to meet the increased requirements, some of the key operational aspects of Google search engines are known.

GOOGLE'S SEARCH ENGINE & PAGERANK

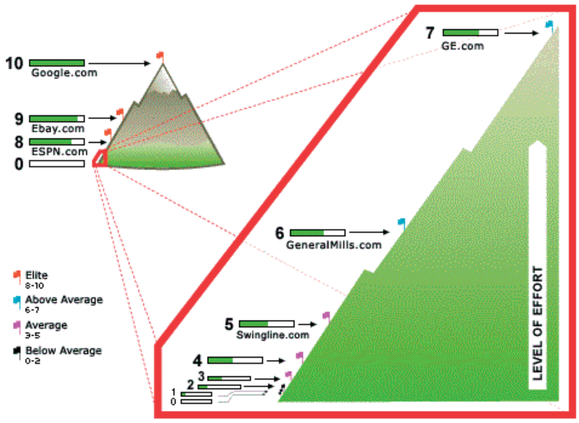
Google search was created by Google founders Larry Page and Sergey Brin in 1997. It originated at Stanford University as part of a research project about a new kind of search engine and resulted in foundation of Google Inc. in 1998 (http://infolab.stanford.edu/~backrub/google.html). In the following years, Google has become the far the most utilized search engine worldwide. In addition to high performance and ease of use, the most valuable factor was the superior quality of search results compared to other search engines. This quality of search results is based on PageRank, a sophisticated methodology of ranking web documents, as well as on variety of other factors that are counted while processing a search query.

Though the algorithms are constantly changing, there are some consistent patterns that stand behind all search results. Firstly, all the links are followed from page to page by Web crawlers (Internet bots that index Web pages). Secondly, the most relevant results appear first. In addition, there are numerous other approaches that vary in their usage (types of search methods, spelling and understanding queries). Despite the fact that Google don't accept payment to increase a site's ranking, they propose sponsored links in addition to the search results..

PageRank is an algorithm of link analysis used by the Google web search engine that assigns a numerical weighting to each element of a hyperlinked set of documents in the World Wide Web. The purpose of this analysis is measuring measuring the relative importance of a page within the set with a premise that the more links are directed to a page, the more important the page is. In other words, the webpages are not considered as parts of website but as individual figures with their own rank. This algorithm determines how trustworthy, reputable, or authoritative a source of the page is. Its main attention is paid to all links between pages to determine their relevance. PageRank (PR) gives a probability distribution to represent the likelihood that a person randomly clicking on links will arrive at any particular page.

PageRank is not only operating upon the total number of inbound links. The basic approach of PR is that a page in fact considered the more important the more other pages link to it, but those inbound links do not count equally. A page ranks high in terms of PageRank, if other high ranking documents link to it. Their rank again is returned by the rank of documents which link to them. Hence, the PageRank of a certain webpage is always determined recursively by the PageRank of other pages. Since the rank of any document in the Web influences the rank of any other, PageRank is, in the end, based on the linking structure of the whole Internet.

Levels of PR vary (picture 2). Each new one is more difficult to reach than previous. Moreover, PR can decrease over time, if the page loses its value and relevance.



Picture. 2. PageRank distribution s function of level of effort (source: http://elliance.com)

Concept and key principle of PageRank

The concept and the algorithm established by the founders of Google as can be mathematically expressed as:

$$PR(A)=(1-d)/N+d[PR(T1)/L(T1)+\ldots PR(Tn)/L(Tn))] \quad (1)$$

Where
PR(A) is the PageRank of the page A
PR (T1-Tn) are PageRanks of the pages linked to A
L (T1-Tn) are outbound links on pages T1-Tn
d is a damping factor ranging from 0 – 1
N is number of the pages assessed

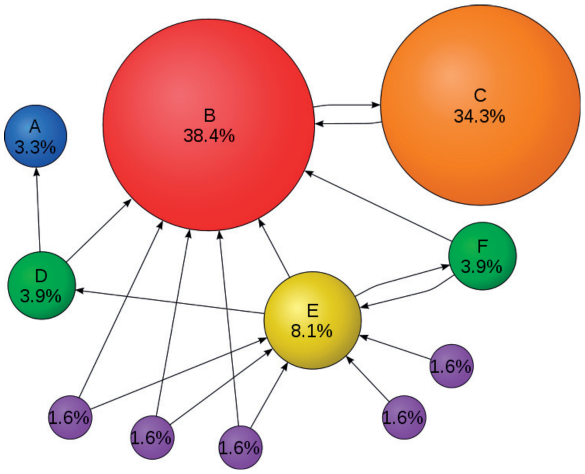
The damping factor is introduced to take into account the chance that a random web user will eventually stop surfing and clicking on the relevant outgoing links. Such behaviour is usually assigned a damping factor of 0.85, although under the specific circumstances it can be assigned other values as well (http://www.personal.psu.edu/users/j/x/jxz203/lin/Lin_pub/2006_ASMBI_1.pdf).

For example, if a page A is linked to pages B with two outbound links, page C with one outbound link and page D with three outbound links, the PR value of the page A will be:

$$PR(A)=(1-d)/3+d[PR(B)/2)+ PR(C)/1)+PR(D)/3)] \quad (2)$$

Usually the initial values given for the individual pages (in this case pages B, C and D) are 0.25. As the spiders go deeper in the site structure, pages linked to B, C and D are evaluated as well and the whole system is recalculated in a number of iterations.

On Page 3 a simple network of pages is presented together with their inbound and outbound links. In this example, page C has a higher PageRank than page E, even though there are fewer links to C. The only link to C comes from an important page and has high value because of that. If a person who starts on a random page has an 85% likelihood of choosing a random link from the page they are currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach page E 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.) Without damping, all web surfers would eventually end up on pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own.



Picture. 3. PageRank and pages network (source: http://en.wikipedia.org/wiki/File:PageRanks-Example.svg)

Other Factors' influence on the PageRank

It has been a subject of many arguments if additional criteria beyond the link structure of the web have been implemented in the PageRank algorithm since its original version. Lawrence Page, one of the creators of PageRank, outlines the following potential influencing factors in his specifications for PageRank:

- Visibility of a link
- Position of a link within a document
- Distance between web pages (achieving fewest number of clicks)
- Importance of a linking page (links from and to high quality related sites are important)
- linking page "up-to-dateness" (how the information corresponds with reality)

In addition, search ranking is determined by other factors that should be considered as well:

- Links from specific sites
- Location (regional)
- Interests – user context (history of web search)
- Number of visits per day
- Key words
- Period of existence of each website
- Time spent by each user
- Trends

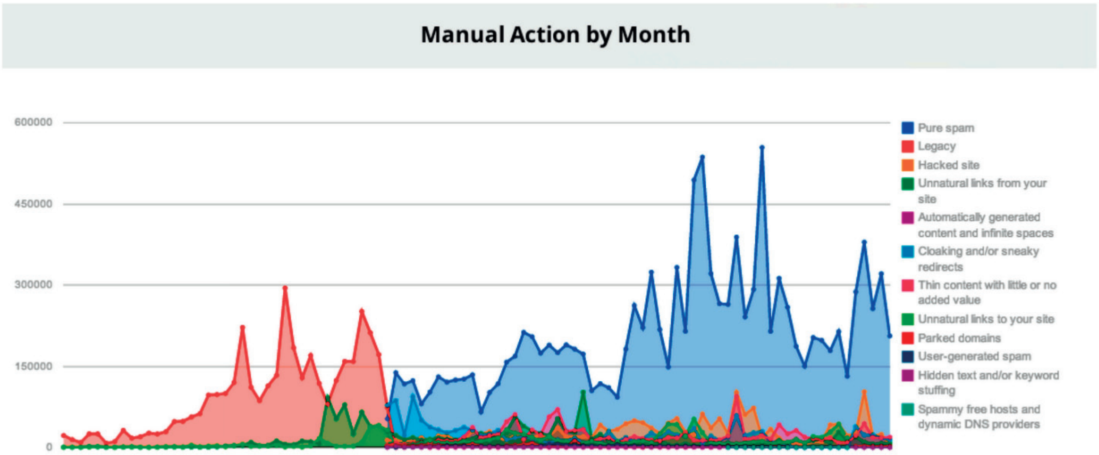
Factors that have no influence on the PAGERANK

Contrary to some popular beliefs, some of the factors do not have influence on the PageRank (PR):

- Frequent content updates do not improve PR automatically. Only relevance is considered in terms of this algorithm
- Content is not taken into account when PageRank is calculated. The goal of PR is to set values to pages according to their connections, not the pages themselves
- High PageRank does NOT guarantee a high search ranking for any particular term.

Filtering the search results

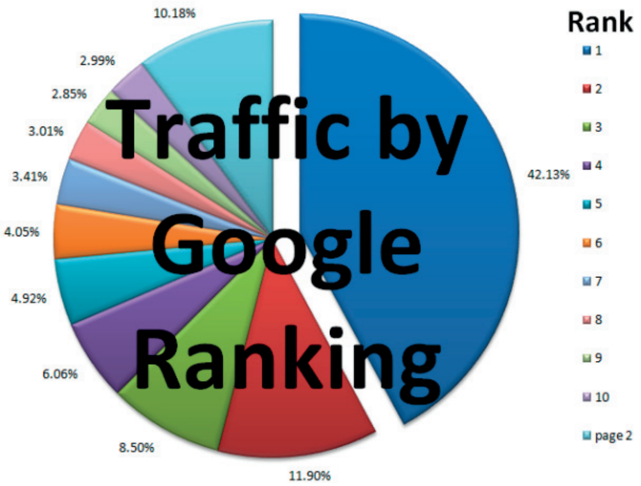
For all of its search results Google applies several methods on filtering and avoiding spam/malware links. For example, it can detect a pattern of unnatural, artificial, deceptive or manipulative outbound or inbound links on the particular webpage. This may be the result of selling links that pass PageRank or participating in link schemes. In some cases pages may contain hidden text and/or keyword stuffing. More than that, they can be hacked by a third party to display spam content or links. Such results are excluded from the final results. Picture 4 shows results of a manual monthly filtering as an attempt to maximize the reliability of the PR results.



Picture. 4. Number of pages monthlyfilterred by Google (source: https://www.google.com/insidesearch/howsearchworks/fighting-spam.html)

PageRank Critisism

Most of criticism and controversy about PageRank is based on the notion that Google has become a virtual "reality interface". Users of the Google search engine tend to ignore all of the search results that do not appear on the top of the list of search results, although some of the valuable information could be found on the lower ranked pages. The AOL analysis from 2006 (Picture 4) show that more than half of the users of Google search click only on the top two Google's search results and that only 10% of the users look for the information which is beyond the first page of the search results.



Picture. 5. Click-through percentages and Google Ranking (source: http://training.seobook.com/google-ranking-value)

Other critics are dealing with the possibilities to influence or misuse the Google algorithms. Some well publicized cases of "Google bombing" (e.g. cases with linking text "miserable failure" or "out-of-touch-executives") have shown that it is possible to have an external influence that can produce unrealistic search results.

In past Google search engine has been also criticized for preferential treatment of Google related sites / pages (e.g. biased ranking for Google Shopping), for personal data collection, for ties with CIA and NSA as well as for susceptibility to political (governmental) pressures. Throughout the years, Google has been continuously improving the algorithm in an attempt to meet the ever-growing demands from the Internet infrastructure, individual web sites / web pages and increased expectations from the web users.

EVALUATING WEB PAGES: CHECKLIST

Once the search engine(s) finish the job, it is still highly recommended to do some basic due diligence and sanity checks in order to confirm that the information retrieved is usable and actionable. The level of such efforts and detail of scrutiny should correspond to the purpose of data collection and risks associated with the use of such data.

Typically, data collected might not be useful if the data source lack competence or integrity. Biased approaches, hidden agendas professional incompetence or other factors might render the search results partly or completely useless. For this reason it is advised to have at hand a check list that covers the following aspects:

- Purpose of the site (information, entertainment, commercial, news, other)
- Site ownership (government, educational, corporate, non-profit, personal, check the DNS extension and WHOIS info)
- Content (quality, quantity),
- Software platform (static, dynamic)
- Usability (information architecture, navigation)
- Look and feel (graphic user interface, usability)
- Maintenance (content updated, obsolete information deleted)
- Security (security systems, legal statements)
- Pages author(s) and posting dates (author's identity, qualification & expertise, date of the information posting)
- Contact information (address, telephone, e-mal)
- Possible vested interest (hidden agenda)
- Overall impression (look & feel)

In most cases, cross-referencing of several reputable sources and consistent information is a good indication of reliability of the information they are delivering.

CONCLUSIONS

Internet is today the most important source of information needed for personal or professional purposes.

Sheer volume of the information makes use of the search engines unavoidable.

Today, Google is by far the most used and the most sophisticated search engine.

Google's continuously evolving complex search algorithm is based on the incoming and outgoing links to specific pages as a main measure of the pages' relevance.

Search results obtained by Google or any other search engine should be considered and used with a degree of skepticism and re-checked to confirm data quality and integrity.

REFERENCES

[1] Google: How Search Works – The Story, http://www.google.com/insidesearch/howsearchworks/thestory/

[2] S. Brin and L. Page, The Anatomy of a Large Scale Hypertextual Web Search Engine, http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf .

[3] eFactory: A Survey of Google's PageRank, http://pr.efactory.de/ .

[4] P. Carven: Google's PageRank Explained, http://www.webworkshop.net/pagerank.html .

[5] GoogleGuide: What's PageRank? http://www.googleguide.com/pagerank.html .

[6] University of Maryland: Web Page Evaluation Checklist, http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/EvalForm_General.pdf

[7] Colorado Stae University: How to Evaluate a Web Page, http://lib.colostate.edu/howto/evalweb2.html#purpose .