Exploring the World Wide Web

István Szűts, Gábor László

Keleti Károly Faculty of Economics, Budapest Tech
Tavaszmező u. 15-17, H-1084 Budapest, Hungary
szuts@bmf.hu, laszlo.gabor@kgk.bmf.hu

Abstract: In this paper, we discuss webometrics. We show its connection to related fields and place it within a complex approach. The paper examines the webometrics as a management tool in different areas.

Keywords: webometrics, scientometrics, webology ,information science

1 Intoduction

The WWW (World Wide Web) was designed in 1989 by Tim Berners-Lee at CERN (the European Organization for Nuclear Research) in Geneva. His proposal discussed the problems of loss of information within complex evolving systems and derived a solution based on a distributed hypertext system. The introduction of the web as a part of the internet in 1991 helped make the internet more popular and easier to use. The web was designed as a "hypertext system" for the purpose of enabling efficient and easy information-sharing among geographically separated teams of researchers.

The web has a greater impact on communications and society than perhaps any other technology. As societies has moved towards becoming "information societies", virtual space has become the communications channel and the tools of socialization. Over the past decade, the importance of electronic communication and the internet has increased significantly.

Some people incorrectly equate the World Wide Web with the internet. Although the web utilizes the internet as its information transmission medium, they are not the same. The web is simply one of the most popular services on the internet.

1.1 Overview of the Architecture of the WWW

The web is the global networked information system of interlinked computer networks that serves files formatted in HTML, XML, PDF, DOC, and other file

types. A document can be static (prepared and stored in advance) or dynamically generated (in response to user input). The files can contain text, images and multimedia components, can include hyperlinks to other such files on different host servers, and can also act as interfaces, in particular, in order to help users to meet their specific information needs.

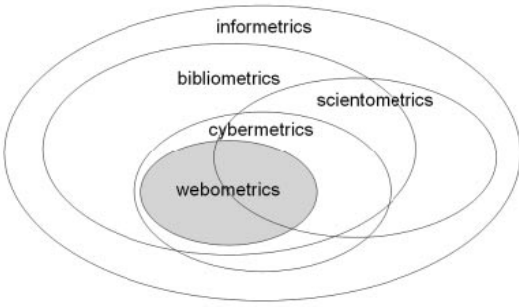
In computing, hypertext is a user interface paradigm for displaying documents which, according to an early definition (Nelson 1970), "branch or perform on request."

Types of webpages could be in general:

- Acedemic, educational
- Business related (B2C, B2B, B2A, etc…)
- Personal

1.1.1 Definitions of Concepts

As a global document network initially developed for scholarly use and now inhabited by a diversity of users, the web has generated a range of new terms for emerging research areas, such as for bibliometrics, scientometrics and informetrics, terms which have been proposed since the mid-1990s.



Source: Lennart Björneborn and Peter Ingwersen: Toward a Basic Framework for Webometrics

Figure 1
Relationships between the library and information science fields of infor-/biblio-/sciento-/cyber-/webo-/metrics

Scientometrics is the science of measuring and analysing science. In practise, scientometrics is often done using bibliometrics, which is the measurement of (scientific) publications. Webometrics and cybermetrics are currently the two most widely adopted terms in Information Science. Webometrics is a field of the information science. The science of webometrics attempts to teake measurements the internet – to get knowledge as to the number and types of hyperlinks, the

structure of the World Wide Web and usage patterns, such as information seeking and search engines. Many of the metrics are subjective, such as peer assessment and selectively.

Albert-László Barabási in his book named *Linked* made an attempt to clarify with scientific methods how everything is connected to everything else and what it means. The social networks on the web are not a new phenomena: for example, there have been Usenet[1] newsgroups. However, the presence of a growing number of social networks such sa iWiW and myVIP in Hungary and orkut run by Google are generating new areas of interest in this phenomena. At the same time, the examination of the web is more than simply a question of science. It could be a business tool, in for example "scanning" the environment as a strategic organizational activity, and researching business information about competitors.

2 Exploring the Web for Business Information

Finding information need not be difficult or time-consuming. The key is to know what one id looking for, and why. With a clear purpose, one can readily decide among the variety of information sources which ones are the most appropriate for the given needs and capabilities.

Multinational companies can, and do, assign teams of researchers in competitive intelligence departments.

2.1.1 Environmental Scanning

"Environmental scanning" is not a new term in management theory and practice. Environmental scanning is research mechanism by which managers discover important events and trends outside their organizations. Scanning the business environment was initially defined as the activity of acquiring information. But it has been redefined with the rise of the web, as the range of research possibilities have extended and the speed of information acquisition has increased. Duncan defines the external business environment as all the factors outside an organization that are taken into consideration by the organization in its decision making. These factors depend on the complexity and dynamism of the environment.

---

[1] Usenet was conceived by Duke University graduate students Tom Truscott and Jim Ellis in 1979. Users read and post email-like messages (called "articles") to a number of distributed newsgroups, categories that resemble bulletin board systems in most respects.

Modes of Scanning

Scanning Modes	Information Need	Information Use	Amount of Targeted Effort	Number of Sources	Tactics
Undirected Viewing	General areas of interest; specific need to be revealed	Serendipitous discovery "Sensing"	Minimal	Many	• Scan broadly a diversity of sources, taking advantage of what's easily accessible • "Touring"
Conditioned Viewing	Able to recognize topics of interest	Increase understanding "Sensemaking"	Low	Few	• Browse in pre-selected sources on pre-specified topics of interest • "Tracking"
Informal Search	Able to formulate queries	Increase knowledge within narrow limits "Learning"	Medium	Few	• Search is focused on an issue or event, but a good-enough search is satisfactory • "Satisficing"
Formal Search	Able to specify targets	Formal use of information for planning, acting "Deciding"	High	Many	• Systematic gathering of information on a target, following some method or procedure • "Retrieving"

Source: http://choo.fis.utoronto.ca/ncb/es/ESmodes.html

Figure 2

Modes of Environment Scanning

2.1.2 Business Intelligence - BI

The term business intelligence (BI) typically refers to a set of business processes for collecting and analyzing business information. Organizations typically gather information in order to assess the business environment, and cover fields such as industry or market research and competitor analysis. Competitive organizations accumulate business intelligence in order to gain sustainable competitive advantage, and may regard such intelligence as a valuable core competence in some instances.

Some people use the term BI interchangeably with "executive information systems" and the information that they contain. In this sense, one can regard a business intelligence system as a decision-support system (DSS).

Business intelligence includes tools in various categories such as marketing and Customer Relationship Management (CRM), Data mining, Management Information Systems (MIS), Knowledge Management (KM), Scorecarding, and so on.

2.1.3 Sources of Information

This part of this article is based on Competitive Intelligence by Industry Canada.

Competitive intelligence is all about analyzing relevant business environment information on an ongoing basis. It includes the analysis of competitors and

suppliers, of technology, industry and market trends, of legal and regulatory changes, and of political and economic changes. To be effective, competitive intelligence needs to become part of the business culture. A strategy needs to be defined, and a plan assembled and implemented. While there are many sources of data, each channel will deliver different value. It is important to define what the goal of the research is before starting, then identify which tools are most likely to deliver that information.

Sources of information on the web or related to web:

- Software

  A variety of specialized competitive-intelligence products are available to automatically search online sources. These have the potential to find information rapidly and easily. Intelligence Agents are a class of software products which are used to automate information capture. These products are customizable online search, retrieval and notification agents. They are sometimes referred to as push technology. Some monitor competitors' websites; some monitor the entire internet for specified information.

- Commercial Services
  - Consulting and Research Services
  - Information Professionals
  - Media Monitoring
  - Training

  Some consulting companies offer advice on competitive-intelligence strategies and methods. Others offer to carry out searches and analyses that cannot be done in-house by the client company. Some companies combine these services, and in some cases offer proprietary software.

- Online Sources

  A wide variety of competitive-information sources are available online. There are websites where one can search directly or by using search engines (such as Google, Yahoo, etc.). Then there are news groups (such as http://groups.google.com/, http://groups.yahoo.com, etc…) which can be commercial news organizations or informal news and discussion groups. Specialised content can be purchased from subscription services, including new filtering services and online databases.

  Websites can give competitive-intelligence researchers immediate access to information about the corporate environment. A company’s homepage is a wealth of information abut that company, more useable information than one might think at first glance.

Many sites are designed as portals, or gateways, to information about particular subjects or groups of subjects. Many portals are multi-functional, offering viewers a chance to shop, participate in discussion groups or play games.

- Offline Sources

  Prior to the advent of electronic media, most competitive intelligence was conducted "offline" through traditional means of research. The major source of primary research is the network of colleagues, customers, suppliers and other contacts. These means are still useful today, and in most cases, very economical. In some cases the traditionally offline sources are freely available online, for example articles of newspapers that have web-based as well as print editions. As well, in many cases the graphs and tables are excluded from the archives of the newspapers so a major source of secondary research is publications, such as newspapers, journals, industry reports, and government reports, which are available through the public library.

The Internet Intelligence Index™ was designed by Fuld & Company Library to help gather information from a wide variety of public services, in support of one its competitive intelligence efforts. It contains links to over 600 intelligence-related internet sites, covering everything from macro-economic data to individual patent and stock quote information.

3 Webometrics in Practice

The web offers the possibility of taking advantage of informal scholarly communication (as was originally proposed by Tim Berners-Lee), an option not available in traditional paper publications. Formal academic and scientific papers are also increasingly being published on the web, maintaining and, indeed, increasing the high standards of peer review, as electronic journals are clearly cheaper than their traditional counterparts and thus more widely disseminated.

A best known running project is the "Webometrics Ranking of World Universities", which is an initiative of the Laboratorio de Internet, Spain.

Webometric indicators are provided to show the commitment of the institutions to online publication and to the worldwide Open Access to knowledge. Evaluation has had a significant impact on the scientific performance of universities and research centres worldwide, prompting researchers to increase the publication of their results and using for that more visible journals. The key actor for this phenomenal success was the (almost) universal adoption of the citation databases as the basis for such evaluation.

3.1 Data Collection

For building the list of universities, several global sources, including some specialized portals (such as 'Universities Worldwide' and 'All Universities around the World') were checked. The unit for analysis is the institutional domain, so only universities and research centres with a clearly identified domain are considered. Currently the project has analysed about 10,500 university domains and 4,000 research related organisations.

The data is collected automatically from the main search engines (currently Google, Yahoo! Search, MSN Search and Teoma are used), using ad-hoc scripts or the APIs provided by the commercial engines.

From a quantitative point of view, three different indicators are calculated from search engines:

- Size

  The number of pages is calculated using four search engines: Google, Yahoo, MSN and Teoma. For each search engine, results are normalised to 1 for the highest value. Then for each domain, maximum and minimum results are excluded and every institution is assigned a rank according to the combined sum.

- Visibility

  The total number of unique external links received (inlinks) by a site can be confidently obtained from Yahoo and MSN only. For each engine, results are normalised to 1 for the highest value and then combined to generate the rank.

- Rich Files

  After evaluating the "academic" relevance and the volume of different file formats, the researchers consider for their purposes the following 'rich files': *.pdf - Adobe Acrobat PDF, *.ps - Adobe Postscript, *.doc - Microsoft Word, *.ppt - Microsoft Powerpoint.

The three ranks were combined according to a formula where each one has a different weight: WR=2S+4V+R (Where WR= Webometrics Rank (Position), S=Size, V=Visibility R=Rich Files)

3.2 Ranking

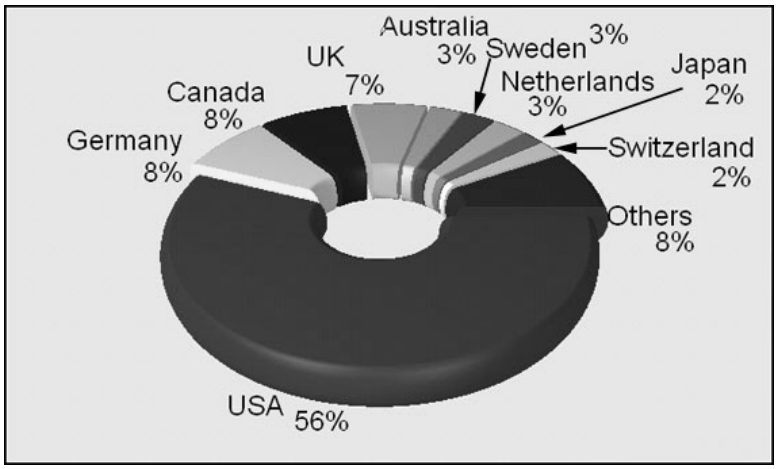
The ranking is based on a combined indicator that takes into consideration the volume of the published material on the web, and the *visibility* and *impact* of these webpages measured by the *"sitations"* (site citations) or links they received (inlinks). It is derived from the Web Impact Factor, built on the same idea as the

bibliographic databases based Impact Factor of the Journal Citation Reports published by the Institute of Scientific Information.

Productivity. The universities are classified by a mathematical combination of the rankings according to their websize, number of rich files, number of papers published in the last ten years and records in the Google Scholar.

Visibility. The criteria combined include link visibility, number of citations to papers in the ISI database and number of visits (popularity) to the web domain.

Impact. The list is presented according to their position in the Shanghai's ranking, and the Webometrics, the Times and ESI citation rankings are also provided for comparative purposes.



Source: http://www.webometrics.info

Figure 3
Top 200 by Country

Although the main purpose of the Webometrics Ranking is to promote publications on the web by universities and other research related institutions, the web indicators produced allow a comparative analysis with other scientometric or bibliometric indicators.

Conclusions

The future of web research promises challenges and opportunities, ones that can be most successfully faced with a multi-disciplinary approach in which information scientists and webometrics can play an important role. We showed in our overview, that the World Wide Web is more than a network, it is the scene of socializing, of business transactions, and of communications, and it is the cornerstone of the information society.

References

[1] Berners-Lee, T: Information Management: A Proposal Retrived from: http://www.w3.org/History/1989/proposal.html

[2] Björneborn, L., Ingwersen, P.: Perspectives of webometrics. Scientometrics, 50(1), 2001, pp. 65-82

[3] Björneborn, L., Ingwersen, P.: Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology, 2004, 55(14): 1216-1227. Full text available: http://www.db.dk/lb/Bjorneborn & Ingwersen 2004 Toward a basic framework for webometrics.pdf

[4] Competitive Intelligence, Industry Canada Retrived from: http://strategis.gc.ca/epic/internet/inee-ef.nsf/en/h_ee00499e.html

[5] Tounkara, T., Benhamou, P.: Knowledge Management and Scientific Observation Retrived from: http://eric.univ-lyon2.fr/~pkdd2000/Download/WS5_08.pdf

[6] Vaughan, L.: Exploring website features for business information, Scientometrics, 2004, Akadémiai Kiadó, Vol. 61, No. 3 pp. 467-477, On-line access via EISZ

[7] Vaughan, L, Wu, G.: Links to commercial websites as a source of business information Scientometrics, 2004, Akadémiai Kiadó, Vol. 60, No. 3, pp. 487-496, On-line access via EISZ

[8] Webometrics Ranking of World Universities Retrived from: http://www.webometrics.info