Theory and Practice of the Webometrics Discipline

Gafar A. F. Ibrahim
Deanship of Libraries - University of Nyala
Introduction:
Studies in Library and Information Science (LIS) used to focus on static tangible information resources for a long time since the invention of writing in the third millennium BC or possibly earlier by the Sumerians. However, the situation has changed dramatically into study of dynamic intangible resources by the mid 1990s when efforts have been made to investigate the nature and metrics of the web through applying techniques and methodologies of bibliometrics and informetrics in order to deal with the continuous flow of web-based information resources.

Statement of the problem:
Since its invention in the last decade of the twentieth century, the web remained untamed and an immeasurable land of no man. Its uncontrolled nature creates numerous problems in the interpretation of results, for instance, from the automatic creation or replication of links. The loose connection between top-level domain specifications (e.g., .com, .edu, and .org) and their actual content is also a frustrating hurdle. For instance, many “.com” sites contain noncommercial content, although “.com” is the main commercial top-level domain.

The importance:
Webometrics has been integrated in the curriculum of several schools of LIS worldwide except in the Middle East and in the Sudan in particular. Its applications in measuring online intellectual production in science, research and development and patents for industrial assessment, among others, is crucial in reinforcing policies and strategies of industrial and higher education institutions.

The objectives:
Webometrics, the quantitative study of web-related phenomena, emerged from the realisation that methods originally designed for bibliometric analysis of scientific journal article citation patterns could be applied to the web, with commercial search engines providing the raw data. It comprises research from fields other than information science such as communication studies, statistical physics, and computer science. This study aims to introduce the field of webometrics as a contribution to library and information science literature. Its major goal is to investigate components, theories and applications of such field through the literature review methodology.

However, this paper investigates the issue of theory and practice of the webometrics discipline in two sections. The first section provides the theoretical framework which tries to revisit old philosophies and theories
relevant to the realm of virtual knowledge. Such revisiting helps selling old brandy in a new bottle, revive the solid thoughts for building better grounds in order to enable webometrics to be understood, assimilated and further developed. The second section is concerned with the ways in which this field is empirically oriented.

Section I
Background:

To understand the intellectual framework of webometrics, a line should be drawn between both the Internet and the web. Researches look at the Internet as the foundation and access on which the web architecture is based (O'Neill 1998). Moreover, the International Telecommunications Union (ITU) provides that "the Internet is a global network of interconnected networks that applies the Internet protocol" (The ITU 2004).

Early in 1951, the first commercial electronic computer known as UNIVAC I was submitted to the US Bureau of Census. Shortly after computers became available at universities, research projects were conducted to link these computers for information sharing (Sadowsky et al., 2003, p.1). Beginning with interconnection of machines at four universities (including two branches of University of California) in 1969 under the name Advanced Research Projects Agency Network (ARPANET), the Internet has grown by leaps and become network of networks of phenomenal proportions (Carl-Mitchell & Quarterman 1994).

But what is known today as the Internet, as Kent suggests, has its origin in the early 1960s as a project of the US Department of Defence (DOD) (Kent et al. 2003; The ITU 2004, op. cit.). Some studies assume that early-recorded historical background of the social interactions that could be enabled through networking was a series of memos written by J.C.R. Licklider from Massachusetts Institute of Technology (MIT) in August 1962 who proposed the concept of "Galactic Network". Licklider envisioned a globally interconnected set of computers through which everyone could quickly access data and programmes from any site. In spirit, the concept looked like the Internet of today. But recent researchers discovered that the documentalist and information science pioneer Paul Otlet had formulated the idea of "International Documentary Network," earlier in 1893 similar to the Internet of nowadays (Rayward 1994, op. cit.). In October 1962, he became

Universal Automatic Computer I.
the first head of the computer research programme at the Defence Advanced Research Projects Agency (DARPA) (Leiner et al. 2002).

Leonard Kleinrock at MIT developed the theory of packet switching that later became the basis for Internet connections. Lawrence G. Roberts, a professor at MIT, connected two computers via telephone lines in 1965. Roberts became part of DARPA in 1966 and developed the plan for ARPANET. Its specific aim was to give American science and technology a boost in response to the former Soviet Union’s launch of Sputnik, the first artificial satellite and a pre-emptive security facility. DARPA wanted to make a network that was smart enough to spontaneously recover from problems such as power failures and interruptions in communication lines in the case of a nuclear attack. It called its network DARPANET (Wiseman 2000). Another hypothesis claims that the actual Internet finds its origin in Robert Kahn’s idea of open-architecture networking, or ‘Internetting’ (Eck 2002, p.14). His idea was that an open architecture would be able to connect multiple independent networks, each network itself having a different operating system and design. Such as open-architecture network required a new communication protocol, the TCP, which was designed in 1973-74 by Robert Kahn and Vinton G. Cerf\(^2\) and later split into TCP/IP (transmission control protocol/Internet Protocol) in 1978 (Cerf 2005; OECD 2005).

The National Science Foundation (NSF) extended its network in the 1980s to increase international educational and research opportunities. This network used NSF computers for data storage, file transfer, and e-mail systems. NSFnet evolved into what is now known as the Internet. It incorporated the TCP/IP format in the UNIX operating system for data packet transmission (Abraham 1995). This format was used to connect the original ARPANET, with a packet satellite network linked the US and Europe, and with a ground-based packet radio network, into the first Internet (Wiseman 2000, p.5).

There is a growing controversy concerning the question of ownership of the Internet domain name space. There are two competing arguments as to the ownership. The first is that since DARPA, and later the NSF funded and developed the Internet, the US government owns the Internet domain name space. Under this line of reasoning, IANA’s\(^3\) authority to manage the domain name systems (DNS) derives from its position as a government

---

2 Known as the father of the Internet, holds a PhD in computer science from University of California Los Anglos and currently Vice President and Internet pioneer at Google, Inc.

3 Internet Assigned Numbers Authority, a subsidiary of the Information Sciences Institute (ISI) of the University of California founded in 1970s (Albert et al. 1999, op. cit., p.20) currently operated by ICANN.
contractor (Albert et al. 1999). According to ICANN⁴ there is growing legal tension between the expansion of the DNS or country code top-level domains (ccTLDs) into generic top-level domains (gTLDs) such as (.aero,.biz,.coop,.info,.museum,.name and .pro) and the business identifiers or trademarks (MarkMonitor 2011).

The proliferation of LIS publications addressing and researching aspects of the Internet is still a relatively new phenomenon. In fact it is difficult, if not impossible, to find an information science periodical without one article dealing with Internet technology (Greenberg 2003).

On the other hand, the World Wide Web (WWW or simply the ‘Web’) is viewed as a content transmission facility that is defined as:

‘An Internet system that distributes graphical, hyperlinked information, based on the hypertext transfer protocol (HTTP). The web is a global hypertext system providing access to documents written in a script called Hypertext Markup Language (HTML) that allows its contents to be interlinked, locally and remotely. It is designed by Tim Berners-Lee at the European Organization for Nuclear Research (CERN) in Geneva’ (Noruzi, 2004).

---

⁴ Internet Corporation for Assigned Names and Numbers.
According to the above definition, the web is a hyperlinked content of information that is transferred on global basis. Diffusion of hypertext content adds new materials and responsibilities to library and information management. The web is a global virtual publishing environment. The term is often mistakenly used as a synonym for the Internet itself, but the web is actually something that is available over the Internet, just like the e-mail and many other Internet services. Tim Berners-Lee, who is head of the World Wide Web Consortium (W3C) so far, created the first server, browser and editor, the HTML code, the URL address and the HTTP transmission protocol at CERN in 1990. CERN released the web into the public domain in April 30, 1993 (CERN, 2003). But Cerf provides that the Internet has entered the public arena in 1994 (Cerf 2005, op. cit.). Another contributor, Robert Cailliau, with Tim-Berners-Lee to the advancement of the web (Sadowsky et al. 2003, op. cit., p.14).Since the last decade of the twentieth century, schools of LIS have inclined to deal with the web as an information phenomenon. Accordingly, scholars have implemented some LIStools and methods such as citation analysis, journal impact factors, bibliometrics and informetrics in order to study and understand aspects of the web and its technology.

**Literature Review:**

Webometrics is a term coined in print article entitled "Informetric analyses on the World Wide Web: Methodological approaches to "Webometrics," published in the Journal of Documentation volume 53,number (4), by Thomas Almind and Peter Ingwersen in 1997 in Denmark. The article identified the Web as an important source for measuring documents and information. After a short period, information scientists recognised that many powerful web measurements could be conducted using the new powerful advanced search features of the search engines. It is defined within the framework of informetric and bibliometric studies. The term is associated with cybermetrics as a generic subfield. In India, three information specialists contributed to the field through writing an article on "A webometric study on library and information science literature," by S. Gopaiakrishnan at Madras Institute of Technology, R. Ambuje at Madras University Library, and S. Seetherama at the Department of Library and Information Science, University of Madras. It is published in Annals of Library and Information Studies, volume 49, number (1), March, 2002. The objectives of their study aims to examine the nature and pattern of citations in the present day library literature and to analyse how far library and information professionals have
made use of web citations in particular (Gopaiakrishnan, Ambuje, & Seetherama 2002). Bjorneborn and Ingwersen, both professors at the Department of Information Studies at the Royal School of Library and Information Science in Denmark have contributed significantly to this field. Their article "Toward a basic framework for webometrics," which is published in the Journal of American Society for Information Science and Technology, volume 55, issue number (4), December 2004, is a cornerstone in the evolution and development of webometrics (Bjorneborn & Ingwersen, 2004). From China emerged a book on webometrics entitled "Webometrics theory and empirical research," by Zhang Yang and published by Science Publication in 2009. Its conceptual merit is based on the quantitative method and it implies the empirical approach. In the UK, Mike Thelwall, who is a member of the Statistical Cybermetrics Research Group at School of Technology at University of Wolverhampton, published an article entitled "Bibliometrics to webometrics," in the Journal of Information Science, issue number(34), August 2008. He investigated in it the evolution of bibliometrics into webometrics due to contributions of works of Eugene Garfied and his Science Citation Index in 1958 (Thelwall, 2008). Another book by Thelwall entitled "Introduction to Webometrics: Quantitative Web Research for the Social Sciences," published by Morgan Claypool in 2009. It is composed for social scientists to measure aspects of the web and explain how this can be achieved on both small and large scale. It also aimed for LIS researchers in the belief that the knowledge and techniques described will be useful for them to guide and aid other scientists in their research. Furthermore, the techniques and issues are all directly relevant to LIS research problem. Thelwall's theoretical thesis is that the obvious quest for webometrics is to support research into web phenomena. His valuable book focuses on three main webometric techniques: web impact assessment, link analysis, and blog searching (Thelwall, 2009). It is clear from the above reviewed literature that the field is in its early beginning. Researches done on IT reliable still few and need to be reinforced in order to generate a scientific ground. There are also many gaps in the theoretical aspect of the field. In October 2009, a framework service contract for expert support with the production and analysis of R&D policy indicators entitled "The use of webometrics for the analysis of knowledge flows within the European Research Area," is developed by IDEA Consult in Brussels (IDEA, 2009).
Recently there is a growing interests in postgraduate studies on webometrics such as the study of Liv Danman Fugl on "Fundamental Methodologies and Tools for the Employment of Webometric Analyses: a discussion and proposal for improving the foundation of webometrics," a master thesis approved at the Royal School of Library and Information Science in Copenhagen in Denmark (Fugl 2001). It is the first study that focuses on the theoretical foundation of webometrics. Its theme connects it with two theories of citations: the normative theory of citations in which scientists cite in order to give credit where credit is due, and to cite the best sources for their purposes: and, the social construction of citations in which scientists cite to gain political advantage, advance their interests, define their claims against attack and convince others (Small 1998). Fugl has published the above master thesis into a book in the same year. In 2004, Lennart Bjorneborn conducted a webometric study on "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach," a PhD dissertation approved at the Royal School of Library and Information Science in Copenhagen in Denmark (Bjorneborn 2004).

Kim Holmberg conducted a PhD study entitled "Webometric Network Analysis: mapping cooperation and geopolitical connections between local government administration on the web," approved in 2009 at Abo Akademi University, Faculty of Economics and Social Sciences, Department of Information Studies in Finland (Kim 2009). Furthermore, Esteban Romero Frias has conducted a study on "Application of webometric techniques to the study of accounting financial variables," which is approved in 2010 at the Department of Accounting and Finance, Faculty of Economics and Business Studies, University of Granada (Frias 2010). It is not surprising to recognize that the above Faculties of Economics adopt information studies within their curricula if we knew that in 1887 a class of seventeen women and three men met in an unused storeroom over the Chapel at Columbia University, ready to begin a course of instruction in librarianship. Such humble beginning of what was called the School of Library Economy, established by Melvil Dewey, marked the inauguration of the first library school in the United States of America (Mount 1988). From Dewey's notion of adding the word "economy" instead of the word "science", is clear that current LIS derives its intellectual grounds from economics and social sciences and may be influenced by ideas of Adam Smith, August Comte and Emile Durkheim. In fact, LIS is an
interdisciplinary field, having the character of metascience, i.e. science of sciences.

Another PhD study was conducted done in 2004 on "Link Analysis Ranking," by Panayltis Tsaparas at the Department of Computer Science, University of Toronto in Canada. Also Ito Takahiko conducted his PhD in 2007 on "Link Analysis with Kernel Metrics," at the Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology in Japan (Ito 2007).

**Definition of Webometrics:**

In order to understand the nature of the term ‘Webometrics’, it is better to shed light on its definitions and associated terms as follows:

“Webometrics is the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the web, drawing on bibliometric and informetric approaches” (Bjorneborn & Ingwersen 2004, op. cit.).

The above definition also includes two approaches ‘bibliometrics’ and ‘informetrics’ which need to be defined. Thelwall added another definition based on content as follows:

"the study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study” (Thelwall 2009).

This definition incorporates a variety of multidisciplinary studies such as web archiving in which the author of this paper has conducted his PhD on "Strategies for Preservation of Web-based Content: Intellectual property barriers to building the Sudan Web Archive," approved in 2010 at the Department of Library and Information Science at University of Khartoum in Sudan.

Tague-Sutcliffe defines ‘bibliometrics’ as follows:

“the study of quantitative aspects of the production, dissemination and use of recorded information” (Tague-Sutcliffe 1992).

Alan Pritchard, who first used the word "bibliometrics," in print in 1969 in an article entitled "Statistical Bibliography or Bibliometrics" in the December issue of the Journal of Documentation, described it as the "application of mathematics and statistical methods to books and other media of communication." Pritchard explained ‘bibliometrics’ as the "metrology" of the information transfer process and its purpose is analysis and control of the process. Also ‘bibliometrics’ is "the study of measurement of the publication patterns of all forms of written communication and their authors" (Hertzel 2003). It is clear that the term has close connection to
‘bibliography’ which also requires more elaboration. According to the Encyclopedia Britannica, the word ‘bibliography’. in its literal sense is derived from the Greek bibliography (2nd century AD), means the writing of books, and it was so defined in the 17th century. Since the 18th century, it has been used to denote the systematic description and history of books. It is now commonly used in two widely divergent, though basically connected sense: (1) the listing of books, arranged according to some system (in this sense it is called enumerative, systematic, or descriptive bibliography) and (2) the study of books as material subjects; i.e., the study of the material of which books are made and the manner in which they are put together (in this sense commonly called critical bibliography). Hence it is the art or science of the description of books (Encyclopedia Britannica 1976). It is the function of bibliography to provide useful information for the student, on one hand supplying him with information about material for study, and on the other helping him to establish the place of a book (or a piece of writing) in an author’s production and its quality and authenticity as a text for study. However, the above-mentioned definition of bibliography has acquired special importance in the 20th century because of the need for effective organization of the records of human communication in the face of the enormous growth of publishing activity and the need, especially in underdeveloped countries, for informed access to the world’s scientific and technical information. It has been said that without bibliography, the records of civilization would be an uncharted chaos of miscellaneous contributions to knowledge, unorganized and inapplicable to human needs.

Additionally, ‘informetrics’ is defined as:

“the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists” (ibid.).

Reference to Thelwall and Vaughan, the webometrics discipline is a new field encompassing contributions from information science, computer science, and statistical physics and its methodology draws especially from bibliometrics (Thelwall & Vaughan 2004). Greer, Grover and Fowler emphasise that webometrics deserves special attention because it gives birth to an entirely new field of specialization within the domain of bibliometrics (Greer, Grover & Fowler 2007). Another dimension, according to studies of Almind and Ingwersen; Bjørneborn and Ingwersen, in the development of webometrics is the application of informetric methods to the web. This new emerging idea of carrying out the same types of informetric analyses on the
web is possible via citation database. It is obvious that informetric methods using word counts can be applied on the web. But what is new is to regard the web as a dynamic citation network where the traditional information entities and citations are replaced by web pages with hyperlinks acting rather like citations (Almind & Ingwersen 1997; Ingwersen 1998; Bjorneborn & Ingwersen 2001).
The philosophy of webometrics aims to quantify the information seeking (web search) behaviour on the web. According to Ingwersen and Bjorneborn, the clearest need for webometrics is to support research into web phenomena (Ingwersen & Bjorneborn 2004). Moreover, webometrics is sometimes used interchangeably with “cybermetrics” which is defined as:

“The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the whole Internet, drawing on bibliometric and informetric approaches (Bjorneborn 2004, p.13).”

Cybermetrics is also a title of an electronic journal that disseminates results from quantitative analysis of the Internet. Its full title is "International Journal of Scientometrics, Informetrics and Bibliometrics," launched in 1997 and published in Spain. The core scientific journal of the field is scientometrics, which started in 1977 in Hungary. The International Society for Scientometrics and Informetrics (ISSI) was established in 1993 and its biannual conferences are one of the main platforms for the presentation of progress and developments in the field since its biannual conference starting in 1996 (Wormell 2001, op. cit.).

Thus, cybermetrics encompasses statistical researches in discussion groups, mailing lists, and other computer-mediated communication on the Internet including the web. In addition to computer-mediated communication using the Internet applications, the above definition also covers quantitative measures of the Internet backbone technology, topology and traffic. It is obvious that the breadth of coverage of cybermetrics and webometrics implies greater overlaps with proliferating computer-science-based approaches in analyses of web contents, link structures, web usage, and web technologies (Bjorneborn 2004, op. cit.). Recently, the overlap between the two terms has been resolved through diverting cybermetrics to be more general and to refer to non-web search such as email or newsgroup studies. However, there are some non familiar terms associated with webometrics such as "librametry," and "technometrics" (Eom 2009).
In accordance with the above mentioned definitions, it is clear that webometrics rely on the quantitative approach but lacks native theory to be better understood. It is a method-centred discipline which has developed various tools for data collection and analysis on the web environment. Consequently, its theoretical grounds are based on citation analysis instead of being created originally for web data. It is noteworthy that the philosophies and theories of citation analysis are derived from the empirical epistemology which may refer to John Dewey though the American pragmatism and capitalism are reflected in almost all aspects of the Internet technology for the Internet is the tool of globalisation (Thelwall 2012). One can infer that the webometric approach is metaphysical since it tries to interpret and describe the nature of reality on a virtual unreal universe.

The Rise of Webometrics:

Historically, the development of webometrics is traced to the first half of the twentieth century from statistical studies of bibliographies and scientific journals (Hertzel 2003, op. cit.). The discipline of bibliography itself goes back to Ptolemy Philadelphus (ca. 285-247BC) who founded the enumerative bibliography as a discipline in the fourth century BC at the Bibliotheca Alexandria and the first bibliographer of notes was Callimachus of Cyrene (ca. 310-240 BC) who might had succeeded Zenodotus of Ephesus in the position of chief librarian (Greetham 1994).

Early studies have uncovered bibliometric power laws such as ‘Lotka’s law’ on productivity distribution among scientists (Lotka 1926); ‘Bradford’s law’ on the scattering of literature on a particular topic over different journals (Bradford 1985); and ‘Zipf’s law’ of word frequencies in texts (Zipf 1935). 

5 Alfred J. Lotka was a mathematician, supervisor of mathematical research in the Statistical Bureau of the Metropolitan Life Insurance Company from 1924 to 1933. It was during this time, 1926, that his definitive work, later called Lotka’s law was produced. His investigation was a productivity analysis. Counting names and the number of publications listed of each, the coverage was for only ‘A’ and ‘B’ names in Chemical Abstracts for 1907 to 1916 and for Auerbach’s Geschichtsafeln der Physik from its beginning through 1900. The data were tabulated and plotted, from which Lotka developed a “general formula for the relation … between the frequency ‘y’ of persons making ‘x’ contributions as ‘x’y = const. Finding the value of the constant when ‘n’ = 2, he observed that: the number [of] persons making ‘2’ contributions is about one-fourth of those making one; the number making ‘3’ contributions is about one-ninth, etc.; the number making ‘n’ contributions is about 1/n of those making one, and the proportion, of all contributors, that make a single contribution, is about 60 per cent. [Notice that Lotka’s observation deals with the least number of productions].

6 This law is formed as follows: R(n) = j log (nl + 1) for the verbal formulation; R(n) = k log nls for the graphical formulation. The constants of these equations are not equal; i.e. j ≠ k and t ≠ s. R(n) is the cumulative total of papers published by the first n journals; n is the rank of a set of journals contributing papers on a particular topic; j, t, k, & s are constants.
Similarly, other power-law distributions have been identified on the web, e.g., the distribution of top level domains (TLDs) on certain subject (Rousseau 1997) or in links per web site (Albert, Jeong & Barabasi 1999; Adamic & Huberman; 2000; 2001). Another crucial development of bibliometrics and scientometrics have been marked by the introduction of citation indexes of scientific literature by Garfield in 1955 which enabled analyses of citation networks in science. In his study “Citations and links as a measure of effectiveness of online LIS journals” Smith has described ‘webometrics’ as synonym of ‘cybermetrics’ and as sub-discipline of bibliometrics (Smith (2004). Access to online citation databases produced a wide range of citation studies such as mapping scientific domains including growth, diffusion, specialisation, collaboration, impact and obsolescence of literature and concepts (Bjorneborn 2004).

Trends towards investigating online citation analysis parallel the recent avalanche of webometric studies. Especially, the vague resemblance between citation networks and the hypertextual inter-document structures of the web gained much interest since the mid-1990s in studies by Bossy (1995); Moulthrop & Kaplan (1995); McKieman (1996); Kuster (1996); Pitkow & Pirolli (1997); Spertus (1997); Ingwersen (1998) in addition to central bibliometric measures of co-citation studies by Small (1973), and bibliographic coupling by Kessler (1963) which have been applied to studies of web clustering, web growth and web searching in studies by Larson (1996); Weiss et al. (1996); Pitkow & Pirolli (1997); Efe et al. (2000); Ding et al. (2002); and Menczer (2002).

Webometrics thus provides potentials for tracking aspects of scientific endeavour traditionally more hidden from bibliometric or scientometric studies such as the use of research results in teaching and by the general public (Bjorneborn & Ingwersen 2001; Cronin 2001; Thelwall & Wilkinson 2003). Accordingly, new terms for describing the emerging discipline have been proposed such as ‘netometrics’ by Bossy (1995, op. cit.); ‘webometry’ by Abraham (1996); ‘internetometrics’ by Almind & Ingwersen (1996).

Section II
Scope of webometrics:
The webometrics discipline is concerned with measuring web-based phenomena such as web sites, web pages, parts of web pages, election web sites, academic web sites, blogs, social networking, words in web pages, hyperlinks, web search engine results and national web domains (Tolosa et
Given the vastness of and easy access to sources of information, there are limitless possibilities for measuring or counting on a huge scale (e.g., the number of web sites, web pages, and blogs), or on small scale, (e.g., the number of web sites in the Sudan, the number of web pages at the University of Khartoum's web site, or the number of blogs mentioning Yassir Arman before the 2010 Sudan Presidential Campaign. While applications of webometrics include range of recent developments, such as patent analysis, national research evaluation exercises visualisation techniques, new applications, online citation indexes, and the creation of digital libraries (Thelwall 2008, op. cit.). Nowadays, webometrics research has expanded from general or academic web analyses to investigations of social websites of blogs, RSS feeds, and study aspects of social networks such as facebook, YouTube, and twitter (Thelwall 2012).

The most powerful webometric method is the web impact factor (WIF) which is proposed in 1998 by Peter Ingwersen. He proposed that a calculation of WIF as a quantitative measure can give indication of the relative attractiveness of countries or research sites on the web at a given point in time. The WIF'S calculation is based on the same concepts as already employed for journal impact factor (JIF) calculation (Ingwersen 1998).

Since the invention of the web there are no serious studies on web-based content aimed to provide statistics, analysis, description and coverage. The Online Computerised Library Centre (OCLC) has launched an annual project known as ‘Web Characterization Project’in order to measure web content and address questions of general interest, such as ‘how big is the Web?’ and ‘what kinds of information are available on the Web?’ and more in-depth, library-oriented issues, such as the rate of migration of print materials to web-accessible formats and the dynamic properties of web resources. It conducted an annual web sample between 1998 and 2002 which analysed trends in the size of content of the web. Its analysis based on the sample that was publicly available. Its statistics were categorised into size and growth statistics, country and language statistics and linkage patterns. Its linguistic statistics show that the majority (72%) of the content on the web is in English language, 7% is in German, 6% is in Japanese, 3% is in both Spanish and French, 2% is in Italian, Dutch and Chinese, and only 1% is in Korean, Portuguese, Russian and Polish respectively and it is obvious that Arabic is less than 1%. The project provides size and growth statistics of the number of web sites as follows: 1998 (2,851,000), 1999 (4,882,000), 2000
(7,399,000), 2001 (8,745,000), and in 2002 (9,040,000). It categorises web site types as public, which provides free, unrestricted access to all or at least a significant portion of its content; private, on which site content is intended for a restricted audience (content need fee payment of authorisation); and provisional, which means site is in transitory or unfinished state (e.g. "under construction") and provides content that is, from a general perspective, meaningless or trivial. Unfortunately, the project did not continue since 2002 (OCLC Web Characterisation Website; O’Neill 1998, op. cit.).

However, the webometrics definition stated earlier, implies quantitative aspects of both the construction and usage aspects of the web providing five main areas of current webometric research including web page content analysis; web link structure analysis; web usage analysis (e.g., log files of users’ searching and browsing behavior); web technology analysis including search engine performance and web ranking.

1) Web page content analysis

OCLC provides two different definitions for the term ‘web site’. One of these definitions rely strictly on physical (network infrastructure) criteria, and the other on information (content-oriented) criteria:

‘Web site (Physical Definition): the set of web pages located at one Internet Provider address.
Web site (Information Definition): a set of related web pages that, in the aggregate, form a composite object of international relevance’ (OCLC 1999).

The web site may contain surface content or deep web content such as databases of open or restricted content. In the extreme, the ‘linked’ nature of the web opens the door for an argument that the web itself is one grand web site.

There has been dramatic change in the use of web resources in academic purposes: if in the past, students and faculty members first referred to academic libraries when they need information, nowadays they rely on and consult the web. This users’ attitude has caused widespread concerns among librarians and information professionals, who fear that the web may replace other sources of information, both printed and digital, available in academic libraries. Such concern derives foremost from the fact that the information available on the web is still only partial and incredible, while much information still exists only in print format. In addition, search engines index only a small amount of the information available on the web and most users do not access the information existing on the ‘invisible web’ which is
considered some 500 times larger than the ‘surface web’. Furthermore, quality of content published on the web is often criticised for being unreliable and lack credibility to be utilised for decision-making, education and research and development. This is because there are no quality assurance or governance policies in the web publishing. Beno argues that while the ‘invisible web’ contains high quality information, we find that the ‘surface web’ is mostly inappropriate for educational or academic purposes and some of it may even be incorrect or biased. Consequently, excessive reliance on the ‘surface web’ may generate superficial research habits, endanger the value of academic information, and adversely affect the quality of research and academic publications (Beno 2009). The International Organisation for Standardisation (ISO8402:1994 Quality Management & Quality Assurance- Vocabulary) dealt with quality assurance as: "those planned and organised activities required providing confidence of a product or service" (ISO 8402:1994).

However, the web is composed of sites on intranets behind firewalls, fee per access sites, and other instances that require prior authorisation. Each site has a unique 32-bit identifier as its IP address. This address, in turn, is divided into 4 octets of 8 bits each often shown separated by dots for instance (132.174.1.5). Since each octet is 8 bits, it can range in value from 8 to 255, creating more than 4 billion potential addresses in the total address space (O’Neil 1998).

Text documents, images, multimedia and many other items of information, referred to as resources or content, are identified by short, unique, global identifiers called Uniform Resource Identifiers (URIs) so that each can be found, accessed and cross-referenced in the simplest possible way. This could be done through breaking down messages into packets by the TCP (Wiseman 2000). The most important feature of the Web is the e-mail service. The first e-mail system had been devised by Ray Tomilson at DARPA, while the first e-mail that had been sent was in 1971 between adjacent computers at DARPA headquarters (Cerf 2005).

According to a UN Report the number of the websites around the world has reached 51,635,284 million on June 2004 with 26.13% compared to 2003. While the number of sites use the Secure Socket Layer (SSL) protocol increased to 56.7% between April 2003 and April 2004 to reach 300,000 web sites (UN Report 2004). Petrov states that web content or web-based
content simply is information, documents, and data published over Internet technology (Petrov 2004).

User-created content (UCC) is a new social impact identified as content made publicly available over the Web (e.g. wikis\(^7\), text blogs, Web 2.0\(^8\), videos, podcasting\(^9\), games, films and music), which reflects certain amounts of creative effort and is created outside of professional routines and practices. Most of UCC activities are undertaken with no expectation of remuneration or profit. Their motives include connecting with peers, self-expression, and achieving a certain level of fame, notoriety or prestige.

2) **Web link structure analysis:**

The idea of link structure and the hypertext goes back when Vanevar Bush proposed in 1945 a machine called a ‘Memix’ that could read and retrieve associative links between different aspects of documents using microfilm. This process was called “associative trails.” But some studies Link the idea of inventing the hypertext and hypermedia systems to the Belgian lawyer and documentalist Paul Otlet (Rayward 1994). However, Ted Nelson (Theodor Nelson) invented the idea of the hypertext in the mid-1960s but the first recorded instance of the word ‘hypertext’ in print only came on February 3, 1965 (Poole2005, p.186).

Henceforth web documents are known as hypertext with links to further associated documents, on the model of references in a scientific paper or cross-references in indexing. With digital documents, these cross-references or nodes can be followed by a mouse-click. Entry to website usually starts at the homepage, which is roughly equivalent to the title page in the print environment. The homepage often provides information about the site, and may also function as a table of contents. Following the homepage, the most essential bibliographic unit on the web is the web page (static or interactive HTML file). Simply, web pages are files lying around on the hundreds of thousands of computers connected to the Internet. The web page is a distinct entity that is identified by a unique URL. Wandering from one document (webpage) to another is called **browsing**. Some people do this just for fun,

---

\(^7\) A website that allows users to add, remove and otherwise edit and change content. An example of this is the online encyclopedia wikipedia that comprised 4.6 million articles in over 200 languages in 2006.

\(^8\) A new application defined as ‘principles and practices that are founded on services rather than software, user control and user participation, scalability from small to large, remixable data sources and data transformations, and the harnessing collective intelligence. These practices include file sharing and social networking such as YouTube, facebook, blogs, RSS, wikis and others. Moreover many of these tools have analogues in the library community. File sharing with MARC is used since 1960s as well as other tools.

\(^9\) A combination of ease of audio production with technologies that allow for subscription and syndication.
following links just to see what’s there. This is usually called ‘surfing the web’ (Stout 1996, p.7; O’Neil 1998). The address of the first Website ever created was ‘info.cern.ch’ and the first Web page address was ‘http://info.cern.ch/hypertext/WWW/TheProject.html’ (CERN Website), while the first library website ever launched was at Virginia Tech University, the University of Michigan and the US Naval Research Library (OCLC 2007).

Sometimes, the web is looked at as ‘digital library’, but finding information on this library is difficult for the quality of content varies significantly as mentioned earlier. As a result, the library and information services have been challenged by the open nature of web publishing (Kuny & Cleveland 1998).

As stated earlier, if we regard the web as a dynamic citation network where the traditional information entities and citations are replaced by web pages with hyperlinks working as citations, it will be clear that link structures have other functions than those of scientific citations and references. Therefore, this new area needs to study such functionalities (e.g., general navigation behaviour in the electronic information environment (Wormell 2001).

3) Web usage analysis:
The broadest interpretation of the web is a collection of HTTP servers operating on TCP/IP interconnected network. Its narrower interpretation, however, consist of all active HTTP servers that receive, understand, and process client requests. Its accessibility can be determined from the response code returned to the client who attempts to get connected. Response codes in the 200-299 range or 300-399 range (the same as DDC classification schema) indicate that the server received, understood, and processed the client request. Response codes in the 400-499 range indicate that a client-side error prevented completion of the request; these include codes that indicate access to the server is not authorised or forbidden. Response codes in 500-599 range indicate that a server-side error prevented completion of the request (O’Neil 1998).

As the Internet grew, there became an increasing need for better tools to access and organise information. Using File Transfer Protocol (FTP) information could be shared between computers. The protocol enables a user to obtain files of text, executable programmes, graphic films, sound files, and other information, but early restrictions of FTP sites allowed only short names with little description about content. Files were often named README, INDEX, NOTICE, ABOUT, and other variations. These file names were thousands of FTP sites across the world, but each site was a collection of files that were not searchable. A user would have to visit each FTP site and
browse through the files to hopefully discover relevant information (Swain 1994).

4) Web technology analysis
Finding information on the Internet seemed somewhat like trying to find a needle in a ‘haystack’. An added dimension to the ‘haystack’ metaphor is that the Internet environment is a dynamic collection of networks. Archie, Veronica, Gopher, Wide Area Information System (WAIS), Mosaic, AltaVisat, HotBot, NorthernLight, Excite, Lycosse, Infoseek, and others were the early search tools of the Internet- the first attempts to provide more order and searchability. Although these early tools are still accessible today, they were developed prior to web browsers and have generally been replaced by more popular web search tools such as Yahoo and Google (Valauskas, 1994; Bjorneborn & Ingwersen, 2001).

5) Web Ranking
A webometrics indicator has been launched in Spain for the purposes of ranking scientific repositories and worldwide universities. Repository ranking indicator provides a list of major research-oriented repositories arranged according to composite index derived from their web presence and the web impact (link visibility) of their contents, data obtained from the major commercial search engines.

The initiative of "Ranking Web of World Universities"of the Cybermetrics Lab, a research group of the Centro de Ciencias Humany Sociales (CCHS), is part of the National Research Council (CSIC), the largest public research body in Spain. It is officially launched in 2004, and it is updated every 6 months (data collected in January and July and published one month later). Web indicators used are based on and correlated with traditional scientometric and bibliometric indicators. The objective of the project is to convince academic and political communities of the importance of web publications not only for dissemination of the academic knowledge but also for measuring scientific production, activities and performance. Moreover, methodology used to create the rakings is based on the unit for analysing the institutional domain. If an institution has more than one domain, two or more entries are used with different addresses. It is found that about 5-10% of the institutions have no independent web presence, most of them located in developing countries. Webometrics is measuring the volume, visibility and impact of the web pages published by universities, with special emphasis in the scientific output (referred papers, conference contributions, pre-prints, monographs, theses, reports,...etc.), abut also taking into
account other materials such as courseware, seminars, or workshops documentation, digital libraries, databases, multimedia, personal pages, ...etc.), and the general information on the institution, their departments, research groups or supporting services and people working or attending courses.

It is recommended in this paper to adopt webometrics for web ranking of other important current trends in scientific research, i.e. measuring scientific production of theses, dissertations and patents. The ProQuest database of theses and dissertations provides a good base for measuring and ranking intellectual productivity by country and themes. The result could assist decision-makers in the governmental agencies to allocate research funding, both in terms of individual research groups and broad initiatives. It is also used to benchmark entire countries' research outputs in order to identify areas of strength and weakness as well as to reveal a pattern of overall improvement or decline. Such work used is done by applying the scientometric approach, while Patent Scope database of the World Intellectual Property Organisation (WIPO) provides the best webometric research setting for measuring and ranking patent applicants and their countries. The later database contains 14,005,945 patent documents including 2,108,759 published international patent applications covered by country.

**Conclusion:**

It is evident that the webometrics as a field of study, methodology and software is emerging worldwide as a strong area of research and study within the LIS. It represents the current trends of the transformation of LIS’s attitude towards regarding the web as an intangible source of information. Its processes can produce improved link counts through counting country code domains or web sites instead of web pages. It is also obvious that its applications range from web citations analyses to ranking and measuring scientific researches and patents productivity. Though some theoretical aspects of the webometric methods are rooted in the empirical approaches, there is a need for further researches on the influences of the pragmatism and capitalism on its intellectual metaphysical nature.

**References:**


