International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

> Homepage > The Journal > Issues Contents > Vol. 15 (2011) > Paper 1

The Journal
Cybermetrics News
Editorial Board
Guide for Authors
Issues Contents
The Seminars
The Source
Scientometrics
Tools
R&D Policy & Resources

VOLUME 15 (2011): ISSUE 1. PAPER 1

An Investigation of Web Resource Distribution in the Field of Information Science

Kun Lu, Soohyung Joo & Dietmar Wolfram

School of Information Studies
University of Wisconsin Milwaukee
P.O. Box 413, Milwaukee, WI 53201 USA
E-mail: **kunlu@uwm.edu**, **sjoo@uwm.edu**, **dwolfram@uwm.edu**

Abstract

This study introduces a new methodology to explore Web information distribution. Subject terms extracted from a key journal in the field of information science were employed to conduct Web searches on Google to identify a corpus of Internet domains and associated Web pages to represent the discipline of information science. A Bradford analysis was then applied to the corpus to determine if the scatter of Web pages conformed to a Bradford distribution. The modeling of the collected data to the power law function in LOTKA program indicates a good fit at even 10% significance level. With a binning procedure and least squares fitting, an R square value of 0.987 was obtained. A division of the data according to top level domain category shows different number of domains and domain productivities in different types of domains. Governmental and commercial domains have higher productivity than the educational and organizational domains. However, the difference between governmental and commercial domains and between educational and organizational domains are not significant.

Keywords

Web information distribution; subject dispersion; domain productivity; power law

Introduction

The World Wide Web (abbreviated as "the Web") has become one of dominant medium for information accumulation, access and retrieval over the past 15 years. The study of the Web to understand its properties and features is an essential prerequisite for its better use. The Web is a dynamic and heterogeneous collection of documents with a fairly loose structure that minimizes the effort to post information and accordingly encourages world-wide participation. On the other hand, its sheer size makes unbiased data collection for research difficult and hinders in-depth understanding of the Web. The similarities and differences between the Web and traditional information media not only inspire researchers to extend traditional methods in Bibliometrics to the Web, but prompts investigators to verify whether the regularities and laws in Bibliometrics are also applicable in the Web environment.

The purpose of this study is to explore whether regularities in information scatter observed by Bradford (1934, 1950) have a similar counterpart in the Web environment. Bradford examined the periodical literature of Applied Geophysics (1928-1931) and Lubrication (1931-June 1933) over several years and noted that if the periodicals are arranged in decreasing order of the number of articles they publish (i.e., productivity) on a given subject and are then grouped as zones that contribute roughly the same number of articles, the numbers of periodicals in each zone will increase as their productivity decreases. Many studies of journal productivity have been carried out on various subject areas since Bradford's time, as outlined by Lockett (1989). More recently, Nicolaisen and Hjørland (2007) have noted that the number of articles contributed by each periodical will depend on how one defines a subject, which can affect data modeling outcomes.

This study contrives a new approach to explore Web information distribution as a method to measure Web resource productivity. Using subject terms collected from a key journal in the field of Information Science, the authors constructed a corpus of Web domains and associated Web pages that could represent Information Science. Then, a Bradford analysis was applied to the corpus to examine whether Web pages within identified domains follow a Bradford distribution in the Web space of Information Science. In addition, the study intends to examine how the productivity of Web resources differs by types of domains.

Literature Review

The present study builds on research from the domain of Cybermetrics and Webometrics, and applies data collection approaches that use Internet search engines to identify and extract relevant data.

Cybermetrics and Webometrics

Many researchers have studied the Web by extending quantitative methodologies in Bibliometrics (Bar-Ilan, 1997; Bar-Ilan, 2000; Molyneux and Williams, 1999; Rousseau, 1997; Thelwall, Vaughan and Björneborn, 2005; Vaughan and Shaw, 2003). Ultimate solutions to some fundamental problems for scientific research are still waiting to be solved (Bar-Ilan, 2001).

Webometrics has been defined as the quantitative studies of Web-related phenomena (Thelwall, Vaughan and Björneborn, 2005). It extends the methodologies originally designed for Bibliometric analysis to the Web environment. Cybermetrics is a broader concept, which includes not only the Web but also some other electronic information resources such as mailing lists, discussion groups and other computer-mediated resources. Although under some circumstances these two terms are used interchangeably, there are some subtle distinctions between them.

Comparative study between Bibliometrics and Webometrics has been an active research topic since the emergence of the latter in the 1990s. Of particular interest has been the study of Web links and their similarity to citations. Rousseau (1997) conducted a "sitation" analysis focusing on the number of in-links to a certain site. The data was found to conform to the Lotka function (Lotka, 1926). Vaughan and Shaw (2003) compared Web citations to bibliographic citations. A significant positive correlation was found between Web citations and bibliographic citations. The results suggested that online and offline citation impacts are in some way similar phenomena. Vaughan and Thelwall (2003) investigated two factors, site age and site content, that influence the creation of links to journal websites. Their findings indicated significant effects for both factors. A more recent study by Gargouri et al., (2010) suggested that online availability will improve citation impact especially for highly cited articles. Other studies have tried to model electronic data to Bradford models. Bar-Ilan (1997) examined how newsgroups reacted to the "mad cow disease" crisis. Data were collected from more than a thousand newsgroups for a period of one hundred days. Bradford's law was shown to be applicable. Faba-Pérez and Guerrero-Bote (2003) examined the distribution of inlinks to websites using Lotka and Bradford models. They found the data did not fit a typical Bradford distribution.

There have been many research studies undertaken to reveal the properties and features of the Web. Albert, Jeong, and Barabasi (1999) investigated the diameter of the Web, which measures the shortest distance between any two nodes. Their findings indicated that despite its huge size, the Web is a highly connected graph with an average diameter of 19 links. Bar-Ilan (2000) conducted a content analysis based on the result pages retrieved from search engines and indicated that the Web has an excellent potential to serve as a source for identifying bibliographic items for research.

Search Engines as a Tool for Metrics Research

The enormous scale of the Web impedes the data collection process for scientific research, which makes it impossible to study the whole Web. Search engines as the major tools for discovering and locating information on the Web have been increasingly used in many academic research studies (Bar-Ilan, 2000; Bar-Ilan, 2004; Bharat and Broder, 1998; Rousseau, 1997; Thelwall, 2002). Research groups closely linked with search engine development are clearly in an advantageous position to provide realistic evidence of the success of their methods (Rasmussen, 2003). More recently, Aguillo et al. (2006) used advanced features of search engines to collect data from Web for university rankings. Cybermetric measures were shown to be useful to reflect the contribution of technological universities.

Although search engines are useful for Internet research, they also have a number of drawbacks. Thelwall, Vaughan and Björneborn (2005) identified four drawbacks of using commercial search engines for scientific research including: their partial coverage, secretive algorithm used for retrieval, unreliable results, and commercial models used for implementing the search environment. The lack of complete coverage of the Web by search engines presents another challenge. Bar-Ilan (2000) noted that it is almost impossible to obtain full coverage of a topic on the Web because of resource discovering and indexing policies used by search engines. Sherman and Price (2001) defined 'Invisible Web' or 'Deep Web' as data freely reachable through the Web but not covered by the search engines. Even search engines may improve their coverage as the technology advances. It appears to be difficult to obtain a complete list of their results (Thelwall, 2008).

A brief review on previous literature has demonstrated that although a number of studies extended Bibliometrics analysis to the Web environment, few have investigated the nature of information scatter among Web information sources. The purpose of this study is to explore the scatter of information on the Web on a topic by using a novel method to conduct a Bradford analysis of Web domains. Using Egghe's (1990) concept of Information Production Processes, classic Bradford analysis of journal productivity consists of journals (sources) that contribute articles (items). This analogy can be extended to the Web where each Web domain, defined by an Internet Protocol address or Uniform Resource Locator (URL), is the equivalent of a source, and each relevant Web page associated with that domain represents an item that contributes to a domain's productivity on the subject of interest. Thus, the domain productivity represents the number of relevant Web pages a given domain produces.

Research Questions

With this framework in mind, the following questions guide the current study:

RQ1 Does the distribution of Web pages on a given topic within specific domains follow Bradford's law of scatter? (What does the distribution of Web resources in a specific topic within specific domains look like?)

RQ2 How does the productivity of a Web domain differ by domain type? A better understanding of these questions is conducive to further exploration of Web information resources. It also has practical implications for Web source evaluation.

Methodology

As is mentioned earlier, it is not feasible to probe the whole Web in Webometric studies. In the present research, the authors selected the search engine Google for identifying available resources on a given topic on the Web. This study acknowledges that what we retrieve from Google is not equivalent to the whole Web. To keep the scope of this exploratory investigation manageable, the topic "Information Science" was selected as the focal subject. To define the scope of Information Science as a subject area, and to reduce the bias inherent in the selection of sources for inclusion (Nicolaisen and Hjørland, 2007), a set of terms that reflect the scope of Information Science was needed. We have assumed that keywords employed in publications in the Information Science field provide a rich set of subject information to describe the breadth of the field. The *Journal of the American Society for Information Science and Technology* (JASIST, **http://www.asis.org/jasist.html**) is not only one of the most highly ranked journals but also contributes a large number of research articles every year. Its scope is broad, it provides good coverage of the field, and it includes subject keywords for its articles.

With this in mind the authors downloaded bibliographic records for one thousand research articles from JASIST for the period of 2001 to 2008 using the Library, Information Science and Technology Abstracts (EBSCO) database. All subject terms from these records were then extracted into a local database. After a process of cleaning and standardization, the frequency of occurrence of each keyword was tallied. Given the long-tailed nature of the distribution of terms, we then chose terms with a frequency larger than three and manually inspected the remaining terms to filter out generic terms (e.g. research). These terms served as the basis for Google searches.

Each of these subject terms was submitted to Google individually using an automated search routine. The returned pages were recorded in a local database for each term. In this study, the top-ranked 100 search results were recorded for each search term. In the case of searches that returned fewer than 100 results, all the retrieved records were included the database. The authors acknowledge the arbitrary cut off at top 100 may lead to biased results. However, given the evidence that users seldom check beyond the third page of results (Spink et al., 2001), which usually represents the top 30 results, it is believed that our sample should be sufficient to cover the important Web pages that most users may concern. In addition, the lower ranked results are believed to be less relevant with the search terms.

The returned pages for each subject term were then combined and checked for duplicates. Two types of duplicate Web pages were found to be significant in the study. The first source of duplicate records represents pages with different URLs but the same content. The second group consists of pages with the same URLs, but that are retrieved by more than one of the subject terms. (e.g. "information science" and "computer science" both have the result page "http://www.cis.upenn.edu/"). In the case of the former, this would not have an impact on our study because pages with the same content are likely to satisfy users' information needs equally. Also, users usually are not likely to mind the source of that page. In the case of the latter, duplicate URLs were counted only once, but the frequency of their occurrence was recorded.

At this point, a list of Web page links dealing with aspects of Information Science has been identified. To investigate the domain productivity, all links were processed to calculate how many Web pages each domain contributed. The overall procedure used for this study is shown in Figure 1.

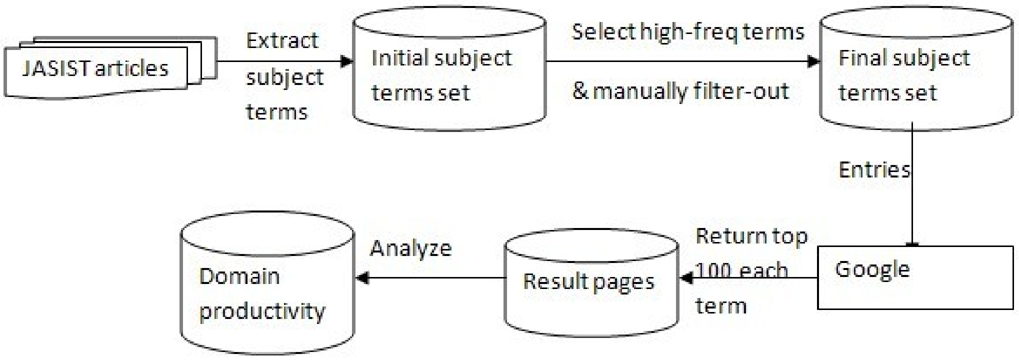

Figure 1: Research Procedure

The products of the analysis were a list of Web domains and the number of pages they contributed. LOTKA (Rousseau and Rousseau, 2000) software was used to fit the data to a power law distribution. The program applied maximum likelihood estimation approach which is preferable to the least square fit for power law distribution. The results of Kolmogorov-Smirnov tests were used as the goodness of fit indicator.

Results

The Distribution of Web resources (RQ1)

In total, 1,388 different subject terms were extracted from the selected 1,000 JASIST articles for the period 2001 to 2008. A strongly inverse relationship between the frequency of a subject term and its rank was found. A standard Zipf equation was used to determine if the observed distribution of subject terms was Zipfian:

$$f(r) = A / r^b \quad (1)$$

where f(r) represents the frequency of the subject term, r represents the rank of that term, and A and b are parameters to be estimated.

The observed data were fitted in Microsoft Excel using the trend line feature against a "power" (Zipfian) model. An R^2 value of 0.9471 was achieved, which indicates a reasonable fit. Figure 2 plots the relationship. As is usually done in power law distribution fits, both axes were represented using logarithmic scaling. Accordingly, a power function will become a line with its exponent as the slope. A logarithmic scaling was used for all of the plots for the data following a power law distribution.

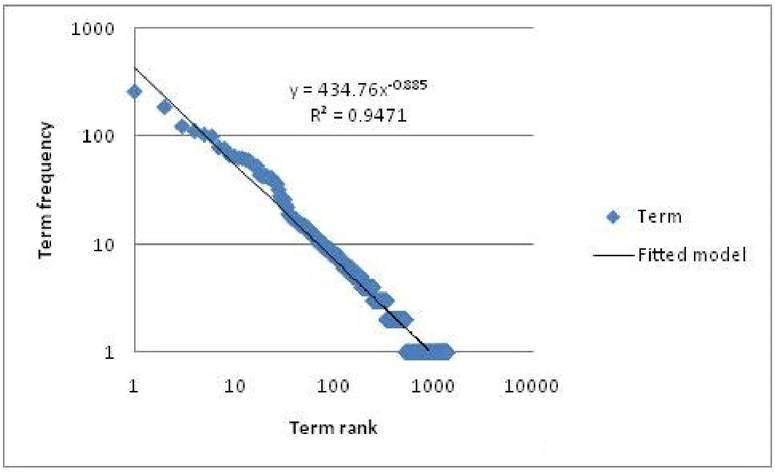

Figure 2: Term frequency and rank for JASIST subject terms 2001-2008

After selecting the subject terms with a frequency of occurrence greater than 3 and manually filtering out meaningless terms (e.g. "research", "books--reviews"), we came up with 273 final search terms. A list of the top 20 search terms is presented in Table 1.

Table 1: Top 20 search terms and the frequency of occurrence

Search term	Frequency
information science	260
information retrieval	187
nonfiction	112
information storage and retrieval systems	104
information technology	100
information services	79
information resources management	77
knowledge management	65
world wide web	63
search engines	63
electronic information resource searching	61
information resources	60
internet searching	53
web sites	53
bibliographical citations	44
web search engines	43
database searching	43
electronic data processing	43
periodicals	41
electronic information resources	41

There were a total of 27,216 Web pages returned by searching each subject term in Google and selecting the top 100 results for each term. After eliminating the duplicate Web pages, as defined above, 26,096 Web pages remained. The domain productivity was calculated by analyzing the URL of each Web page. For example, if a Web page has a URL "http://www.academicinfo.net/infosci.html", it is considered to be an item of the domain "http://www.academicinfo.net/".

There were 14,142 different domains identified. The following table 2 shows the top ten domains with the highest productivity for Information Science:

Table 2: Top ten domains with highest productivity

Domain name	Domain productivity
http://www.amazon.com/	352
http://books.google.com/	342
http://en.wikipedia.org/	311
http://portal.acm.org/	132
http://ieeexplore.ieee.org/	100
http://www.springerlink.com/	94
http://onlinebooks.library.upenn.edu/	81
http://citeseer.ist.psu.edu/	79
http://www.librarything.com/	77
http://www.amazon.co.uk/	73

The domain productivity can be obtained by counting how many Web pages in the dataset came from each domain,. Figure 3 plots the relationship between domain productivity and the number of domains with that productivity. Not surprisingly, an inverse relationship was observed between the number of Web pages per domain in the chosen subject area and the number of domains in which the Web pages appear.

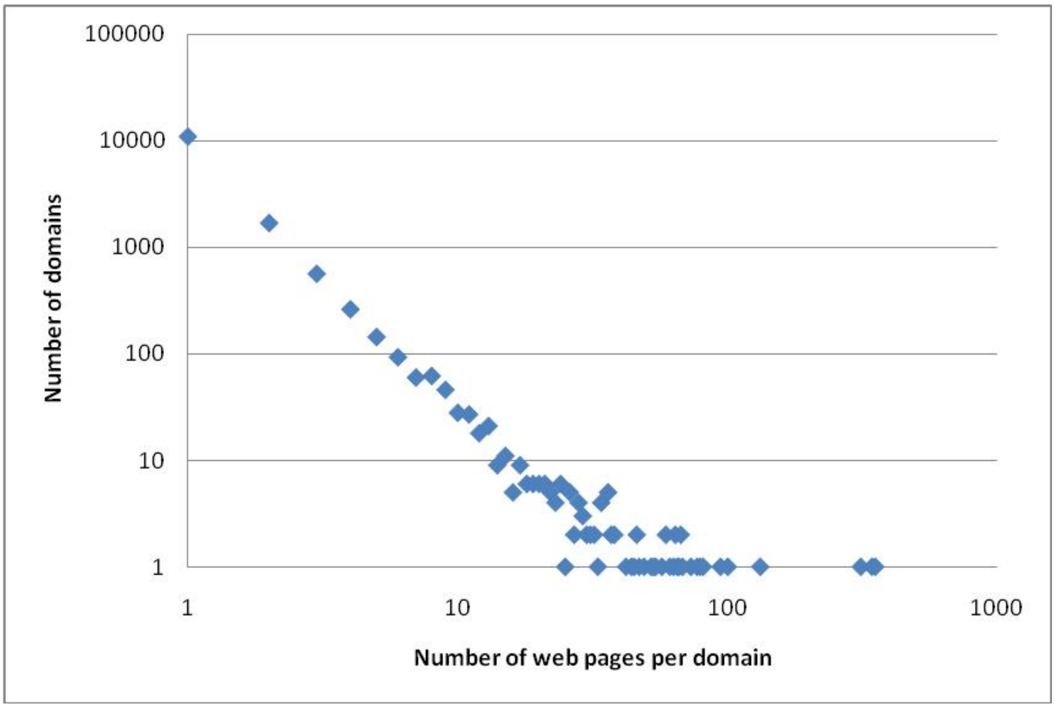

Figure 3: Domain productivity distribution

Earlier research has demonstrated that Lotka, Bradford and Zipf functions are equivalent under certain conditions (Chen & Leimkuhler, 1986). Additionally, it has been suggested that the general Lotka function (for size-frequency data) and the equivalent generalized Leimkuhler function (for cumulative rank-frequency data) provide the best fits for Bradford curves (Rousseau, 1994). Based on this evidence, we decided to fit a size-frequency Lotka function for the domain and Web page Bradford curves using the LOTKA program developed by Rousseau and Rousseau (2000). Specifically, the Lotka function is given as follows:

$$f\,(k) = C / k^{ß} \qquad (2)$$

where *f(k)* represents the number of domains that produce *k* relevant Web pages in the context of this study. *C* and *ß* are two parameters to be estimated for the model. The LOTKA program employs a maximum likelihood approach for parameter estimation and Kolmogorov-Smirnov goodness-of-fit testing for model fitting.

The results produced by LOTKA program indicates a model fit at the 1%, 5% and 10% level of significance. The estimated model is:

$$f\,(k) = (0.767*14142) / k^{2.606} \qquad (3)$$

The fitted line and data were plotted in Figure 4 to give a sense of the estimated model.

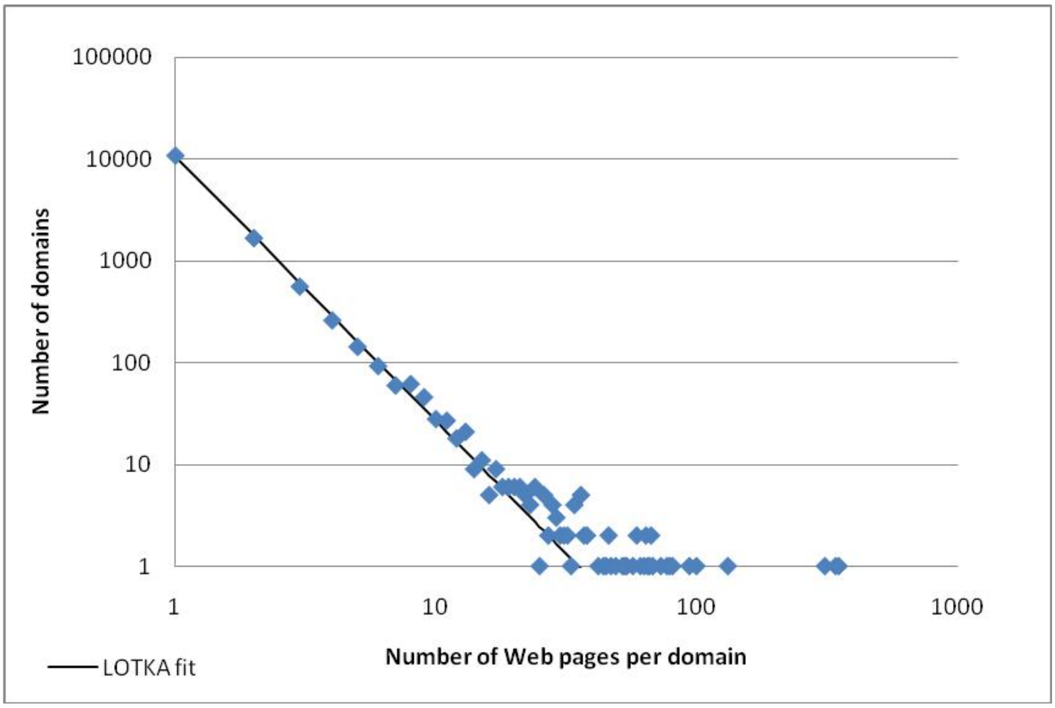

Figure 4: LOTKA fit on domain productivity distribution

A recent study by Milojevic (2010) suggested a partial logarithmic binning procedure to extract the nature of the power law distribution from noisy data. We applied the binning procedure to our data and compared a least squares fitting outcome on the binned data with the previous maximum likelihood fit produced by LOTKA. The result is shown in Figure 5.

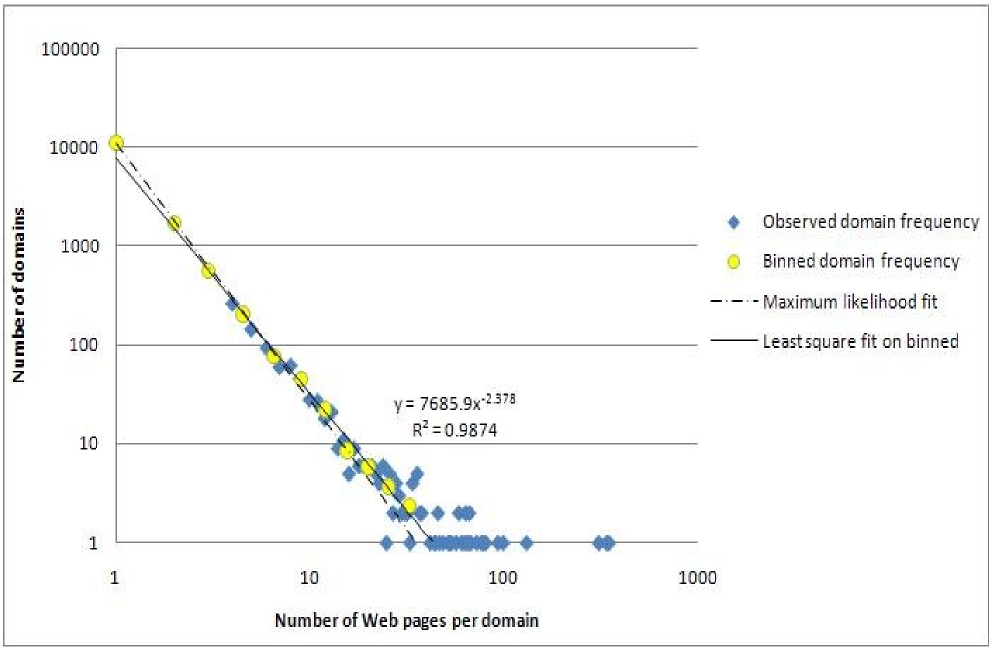

Figure 5: Least square fit on binned data and maximum likelihood fit

After the logarithmic binning procedure, which combines the noisy data at the tail end, the nature of the distribution is more obvious. . The fitting of the binned data using least squares estimation produced an R square value of 0.987, which indicates a reasonable fit. This gives us more evidence to understand the distribution of our domain productivity data. Based on the data collected, it is believed that the distribution of Web pages on a given topic within specific domains follows a power law distribution. Most domains produce few relevant Web pages on the topic while a few prolific domains produce the most relevant Web pages.

The Productivity of Each Web Page by Type of URL Domain (RQ2)

Our second research question focuses on understanding the domain productivity of a given topic (i.e. information science) by different types of domains. The domains were categorized according to the set of top level domain categories (Postel and Reynolds, 1984). The top level domain categories include GOV (government-related domains), EDU (education-related domains), COM (commercial related domains), ORG (organization-related domains) and others. Table 3 tabulates the average domain productivity, their standard deviation and number of domains by domain categories.

Table 3: Domain productivity and number of domains by categories

Domain type	Average Domain Productivity	STD	N
Commercial sites (.com)	1.98	8	5,163
Educational sites (.edu)	1.81	3.08	2,806
Organizational sites (.org)	1.91	7.36	2,838
Governmental sites (.gov)	2.26	4.72	532
others	1.5	2.29	2,803

According to table 3, most of our domains belong to commercial sites, which account for 36.51% of total number of domains, while the governmental category has the smallest number of domains and only accounts for 3.76% of the total number of domains. Although only occupying a small percentage of the domains, governmental domains have a higher average domain productivity than the rest of the categories. A non-parametric Kruskal-Wallis one-way analysis of variance test was applied to test if there is any significant difference between the domain productivity in each category. The test shows a significant difference (d.f.=4, p=0.000) among the five categories in domain productivity. Follow up pairwise comparisons were conducted with Mann-Whitney U tests. Governmental sites have significantly higher productivity than educational sites (d.f.=1, p=0.000) and organizational sites (d.f.=1, p=0.000). Additionally, commercial sites have significantly higher productivity than educational sites (d.f.=1, p=0.000) and organizational sites (d.f.=1, p=0.000). However, there is no significant difference between governmental sites and commercial sites (d.f.=1, p=0.589), or between educational sites and organizational sites (d.f.=1, p=0.618).

Discussion and Conclusions

This study developed a procedure to investigate Web information scatter for a given topic. By relying on a core journal for the topic of interest, subject terms for Web-based searches could be collected to identify Web pages and associated domain data on the distribution of Web pages on the topic of interest across Internet domains. As far as we are aware, this approach to the study of information scatter has not been used before on Internet-based resources. Several findings emerged.

First, in terms of the subject term frequency, we confirmed that the subject terms used in JASIST conform roughly to a Zipf distribution (Power law) with an R^2 goodness-of-fit value of 0.9471. It is noted that these subject terms were assigned to EBSCO databases by professional indexers using a controlled vocabulary (thesaurus). The frequently assigned terms may represent hot topics in the journal.

Second, to verify whether Bradford's law of scatter exists in the Web environment or not, we introduced the idea of domain productivity, which is defined as the number of relevant Web pages retrieved under a given domain name on a specific topic. The results showed there is a strong inverse relationship between the number of Web pages about the subject investigated and the number of domains in which the Web pages appear. Most domains only include a small number of relevant web pages while a small number of prolific domains produce a large portion of relevant web pages. Further modeling of the data with a power law function using the LOTKA program passed at even a 10% significance level with Kolmogorov-Smirnov goodness-of-fit test. An additional logarithmic binning procedure was employed to uncover the true distribution from noisy data. A least squares fit of the binned data yielded an R square value of 0.987. The slope produced by the maximum likelihood estimate was comparable to what was computed by the least square estimate on the binned data. This gives us more evidence to believe the distribution of our data conforms to a power law distribution.

Third, an examination of different categories of domains in our data reveals that commercial sites occupy the largest portion in the number of domains for Information Science, while the government sites account for the smallest portion. This result may reflect the dominance of commercial Web sites on the Internet. Even though the terms were extracted from scholarly journal, the largest portion of retrieved documents is allotted to commercial site category due to the dominance of commercial sites. On the other hand, since the registration of governmental sites is more restricted, the number of domains in that category may also be smaller. Another possible reason may be search engine bias, where Google may index more commercial web sites. A comparison of the domain productivity in each category showed that the governmental and commercial domains have significantly higher productivity than the educational and organizational domains. However, the difference between governmental and commercial domains and between educational and organizational domains is not significant. Therefore, according to our data, the governmental and commercial domains are more prolific in the topic of Information Science.

As with any study, there are limitations inherent in this study that arise from the methodology used. First, there is no practical way to study the whole Web. Even Google, the search engine with the largest index, does not index the complete Web. There may also be bias in the way Google indexes the Web. A meta search engine may be used to replace Google in the current study. However, even a meta search engine doesn't solve the problem completely.

Second, our findings are limited by the scale of this research arising from the subject term frequency cutoff and the limit to the top 100 returned results. It could cause us to ignore some less frequently appearing subject terms and lower ranked search results. We believe that those subject terms are more peripheral to the topic and those lower ranked results are less relevant.

Third, how one defines a subject will affect how productivity is measured, as noted by Nicolaisen and Hjørland (2007). By relying on hundreds of indexer assigned subject terms as a basis for retrieval, the present study has cast a broad net for Information Science-related Web pages. A narrower focus for the field that uses a smaller vocabulary might result in a different distribution of domain productivity.

This study demonstrated that subject dispersion over the indexed Web roughly follows Bradford's law at the Internet domain level, where a domain is comparable to a journal and Web pages within a domain are comparable to journal articles. In the selected topic, commercial and governmental domains have higher domain productivity than educational and organizational domains. Additional research is needed to determine whether this is the case for other disciplines. If future results are found to be similar for other subjects, this may have implications for search engine ranking of retrieved websites. For example, in addition to ranking individual Web pages, domains may be ranked with Bradford relevance zones for further user investigation.

One important contribution of this study is to provide a method to investigate subject dispersion on the Web. As long as a list of subject terms can be identified to represent the topic boundary, the domain productivity figures can be obtained with our proposed approach. Although the approach has limitations, we are not aware of a better approach in current use. Larger scale studies are planned based on the promising findings of this study. The same methodology can be adopted to study other subjects, or can focus on sub-areas within Information Science. A future study will also examine the effect of different weighting systems based on keyword frequency and Web page rank.

References

Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators, **Journal of the American Society for Information Science and Technology**, 57(10), 1296-1302.

Albert, R., Jeong, H., and Barabasi, A. L. (1999). Diameter of the World-Wide Web, **Nature**, 401(6749), 130-131.

Bar-Ilan, J. (1997). The 'Mad Cow Disease', usenet newsgroups and bibliometrics laws, **Scientometrics**, 39(1), 29-55.

Bar-Ilan, J. (2000). The Web as an information source on informetrics? A content analysis, **Journal of the American Society for Information Science**, 51(5), 432-443.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes: A review and analysis, **Scientometrics**, 50(1), 7-32.

Bar-Ilan, J. (2004). The use of Web search engines in information science research, **Annual Review of Information Science and Technology**, 38, 231-288.

Bharat, K., and Broder, M. R. (1998). A technique for measuring the relative size and overlap of public Web search engines, **Computer Networks and ISDN Systems**, 30, 379-388.

Bradford, S. C. (1934). Sources of information on specific subjects, **Engineering**, 137, 8-96.

Bradford, S. C. (1950). *Documentation*. Washington D.C.: Public Affairs Press.

Chen, Y.-S., and Leimkuhler, F.F. (1986). A Relationship Between Lotka's Law, Bradford's Law, and Zipf's Law, **Journal of the American Society for Information Science**,37(5), 307-314.

Egghe, L. (1990). The duality of informetric systems with applications to empirical laws, **Journal of Information Science**, 16, 17-27.

Faba-P&eacute;rez, C., and Guerrero-Bote, V.P. (2003). "Sitation" distributions and Bradford's law in a closed Web space, **Journal of Documentation**, 59(5), 558-580.

Gargouri, Y., Hajjem, C., Lariviere, V., Gingras, Y., Carr, L., Brody, T., et al. (2010). Self-selected or mandated, open access increase citation impact for higher quality research. *PLOS ONE*. <**http://arxiv.org/PS_cache/arxiv/pdf/1001/1001.0361v2.pdf**> (10 December 2010)

Lockett, M.W. (1989).The Bradford distribution: A review of the literature, 1934-1987, **Library and Information Science Research**, 11, 21-36.

Lotka, A. J. (1926). *The frequency distribution of scientific productivity*. Journal of the Washington Academy of Sciences, 16(12), 317-323.

Molyneux, R. E., and Williams, R. V. (1999). Measuring the Internet, **Annual Review of Information Science and Technology**, 34, 287-339.

Milojević, S. (2010). Power Law Distribution in Information Science, **Journal of the American Society for Information Science and Technology**, 61(12), 2417-2425.

Nicolaisen, J. and Hjørland, B. (2007). Practical potentials of Bradford's law: A critical examination of the received view, **Journal of Documentation**, 63(3), 359-377.

Postel, J. and J. Reynolds. (1984). "Domain requirements", **RFC 920** <**http://www.rfc-editor.org/rfc/rfc920.txt**> (15 December 2010)

Rasmussen, E. (2003). Indexing and retrieval for the Web, **Annual Review of Information Science and Technology**, 37, 91-124.

Rousseau, R. (1994). Bradford Curves, **Information Processing and Management**,30(2), 267-277.

Rousseau, R. (1997). Sitations: An exploratory study. **Cybermetrics**, 1(1). <**http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.pdf**> (20 May 2010)

Rousseau, B., and Rousseau, R. (2000). LOTKA: A program to fit a power law distribution to observed frequency data. **Cybermetrics**, 4(1). <**http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html**> (10 January 2010)

Sherman, C., and Price, G. (2001). *The invisible Web*. Medford, NJ: Information Today, Inc.

Spink, A., Wolfram, D., Jansen, M. B. J., and Saracevic, T. (2001). Searching the Web: The Public and Their Queries, **Journal of the American Society for Information Science and Technology**, 52(3), 226-234.

Thelwall, M. (2002). Methodologies for crawler based Web surveys, **Internet Research**, 12(2), 124-138.

Thelwall, M. (2008). Extracting Accurate and Complete Results from Search Engines: Case Study Windows Live. **Journal of the American Society for Information Science and Technology**, 59(1), 38-50.

Thelwall, M., Vaughan, L., and Björneborn, L. (2005). Webometrics. **Annual Review of Information Science and Technology**, 39(1), 81-135.

Vaughan, L., and Shaw, D. (2003). Bibliographic and Web citations: What is the difference? **Journal of the American Society for Information Science and Technology**, 54(14), 1313-1324.

Vaughan, L., and Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? **Journal of the American Society for Information Science and Technology**, 54(1), 29-38.

Received 19/January/2011
Accepted 20/February/2011