

Information Scattering

Suresh K. Bhavnani

Center for Computational Medicine and Biology

Medical School

University of Michigan

&

Concepción S. Wilson

School of Information Systems, Technology and Management

University of New South Wales

Sydney, Australia

Abstract

Information scattering is an often observed phenomenon related to information collections where there are a few sources that have a high density of relevant information about a topic, while most sources have only a few. This chapter discusses the original discovery of the phenomenon, the types of information scattering observed across many different information collections, methods that have been used to analyze the phenomena, explanations for why and how information scattering occurs, and how these results have informed the design of systems and search strategies. The chapter concludes with future challenges related to building computational models to more precisely describe the process of information scatter, and algorithms which help users to gather highly scattered information.

Keywords

Information scatter, bibliometrics, bibliometric laws, Informetrics, Bradford distribution

Introduction

Whether one searches for research papers about a subject in journals, or for facts about a topic on the web, a common observation is that there are very few sources (e.g., journals or webpages) that contain a high density of relevant information, while most sources have only a few. This phenomenon is referred to as *information scattering*, and has been a topic of research for over seven decades.

Although the first report on the phenomenon of information (or literature) scattering is generally attributed to Bradford [1], it became increasingly clear by 1900 that a large part of the scientific literature on any subject seemed to be scattered across an indefinite number of generalist and specialist journals, with no direct relationship to the subject [2]. This phenomenon posed a problem for the bibliographic control of information because of the possibility of “missed literature” in any comprehensive search.

Bradford, a chemist turned librarian, helped to quantify this observation [1]. He recorded a regularity in the numbers of papers on two scientific subjects across different journals. He noticed that if the journals in each of the two scientific subjects were first ranked in order of the numbers of papers they contain on a subject, and then divided into, say, three groups with an equal number of papers, the numbers of journals in successive groups grew in the ratio of $1 : k : k^2$. For example, if 3000 papers on a subject

were found in 800 journals, then the ratio of (ranked) journals, in successive groups of 1000 papers, is 42 : 158 : 600, or approximately 1 : 3.8 : 3.8² [3]. The number of pre-selected groups can be other than three, which will result in a different number of journals in each group, but will result in a similar regularity. This mathematical regularity, which predicts a core, middle, and peripheral groups of journals for any subject, is often referred to as *Bradford's Law of Scattering* and the resulting distribution of papers across journals as a *Bradford Distribution*.

When Bradford published the above quantitative findings, they were surprising as it suggested that the “missed literature” could be as much as two-thirds of the total for any subject. However, these results are relevant even today because current electronic databases (e.g., Web of Science or Scopus) analyze only a fraction of the world’s journals leading to a similar result.

Subsequent to the publication of Bradford’s [1] paper, there has been considerable debate about inconsistencies in the verbal and graphical description of the law [4], and how to represent the law of scattering using more standard statistical measures such as frequency distributions. The latter has led to comparisons [5] with other similar laws such as Zipf’s Law [6] and Lotka’s Law [7] which describe skewed distributions for other kinds of information phenomena (e.g., the distribution of words across books or authors across papers). Such hyperbolic distributions are referred to as the informetric laws and are studied in the broader field of Informetrics, generally defined to include all quantifiable aspects of Information Science [3, 8]. Bibliometrics, including bibliometric laws are (earlier) terms still used in the literature of information scattering.

Furthermore, there have been debates on how to define the concept of “subject” [9] or “topic” [10] (e.g., melanoma versus melanoma treatment), which could produce different results when deriving Bradford distributions of papers across journals or studying the distributions of authors across papers (Lotka’s Law). However, despite these debates about specific details on the original formulation of Bradford’s law of scattering, there is now general agreement that information scattering is a fundamental information phenomenon, and continues to be an active research area for bibliometricians, informetricians, scientometricians, webometricians and other researchers.

This chapter provides an overview of the research on information scattering based on five sections: (1) Types of information scatter; (2) Methods to analyze information scatter; (3) Explanations for the process of information scatter; (4) Implications of information scatter for search strategies and the design of search systems; and (5) Future research challenges.

Types of Information Scatter

One way to classify the different types of information scatter is based on how different granularities of *information objects* (e.g., papers, words, facts) are distributed across different granularities of *containers* (e.g., journals, books, webpages). The different studies on information scatter have analyzed different combinations of information objects and containers, at different granularities. As discussed in the introduction, Bradford [1] analyzed the distribution of papers across journals, and Zipf [6] analyzed the distribution of words within a book. More recently, research has analyzed the distribution of articles across online databases [11, 12, 13], the distribution of images across databases [14], and the distribution of facts about a topic across webpages and websites [15, 16]. In each case, while the exact fitted curves (e.g., power law, exponential, poisson) of the distributions vary, the researchers have found strong similarity to the overall regularity originally observed by Bradford. For example, as shown in Figure 1, the distribution of facts across websites is best fitted by a discrete exponential curve [15].

Nicolaisen and Hjørland [9] have also classified the different types of information scatter in terms of (1) lexical scattering (words across collection of texts), semantic scattering (concepts across texts) and subject scattering (items useful to a given task or problem). Furthermore, the different types of information scatter constrain the relationships between object and container. For example, facts and articles about a topic can be in multiple webpages or databases respectively. However, a particular article can be in only one journal. These constraints result in different relationships between information objects and containers.

Methods Used to Analyze Information Scatter

There have been three principle methods used to quantitatively characterize information scatter: (1) Frequency distributions, (2) Coverage analyses, and (3) Network visualizations and analysis.

Frequency Distributions

Frequency distribution is a standard statistical method designed to show for a dataset the relationship between a ranked list of categories of observations (e.g., 1-10 journals, 11-20 journals, etc.), and the number or frequency of data items that fit into each category (e.g., 200 articles occurring in 1-10 journals). Typically the categories are placed on the x-axis and the number or frequency is on the y-axis, however, several researchers such as Zipf [6] chose to transpose them. Figure 1 shows the frequency distribution of the number of facts about a topic, across a ranked list of pages [15]. Depending on the data and point that the researcher wishes to make, frequency distributions have been used in formats other than the above to characterize information scatter. For example, while the graph in Figure 1 represents unique occurrences of facts in each category, a *cumulative frequency* distribution includes successive values. Furthermore, when there are a large number of categories and frequencies, the distribution are often plotted on a log-log plot which significantly reduces the length of the x-axis for such large datasets, while preserving the overall relationship.

Coverage Analyses

While the above distributions describe the data, they cannot reveal the minimum number of information containers to visit to get all the information objects. To reveal such details, researchers use distributions

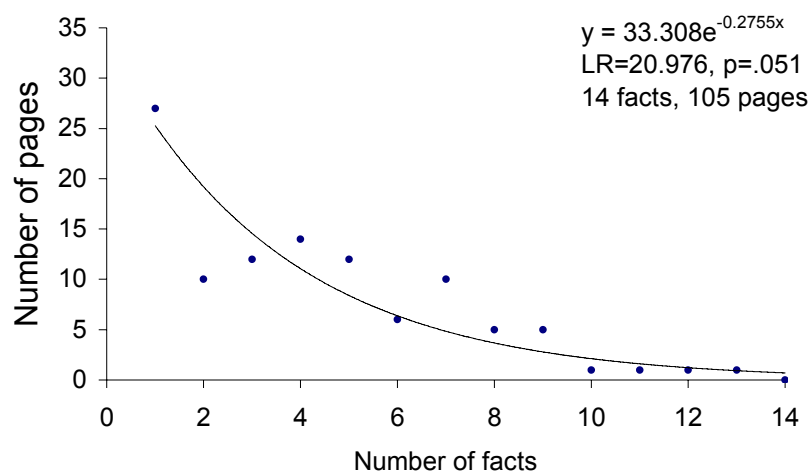


Figure 1. Distribution of facts about melanoma risk and prevention across high quality healthcare pages (Bhavnani, 2005). The distribution is best fit by a discrete exponential curve. Reprinted with permission of John Wiley & Sons, Inc.

which are focused on coverage. Figure 2 shows how many databases are needed to find all the papers related to a subject [13]. Such a distribution provides another measure for comparing the information scatter among topics or databases. Similar to Figure 2, most results in information scatter studies have revealed that it takes many containers to get full coverage of information objects.

Network Visualizations and Analyses

Graphical networks are increasingly being used in a wide range of domains to analyze complex relationships such as information scatter. A network is a graph consisting of nodes and edges; nodes represent one or more types of entities (e.g., facts or webpages), and edges between the nodes represent a specific relationship (e.g., a webpage contains a fact). Figure 3 shows a bipartite network (where edges exist only between two different types of entities) of how facts about melanoma risk and prevention, are contained in webpages from the top 10 healthcare websites for melanoma information.

Networks have two advantages for analyzing information scatter. (1) They represent a particular relationship between different nodes and therefore can reveal patterns, such as how specific facts occur in different groups of webpages. (2) They can be rapidly visualized and analyzed using a toolbox of network analysis methods and visualization algorithms. For example, Figure 3 shows a force-directed layout algorithm which helps to visualize the relationship between webpages and facts [17]. The algorithm simulates placing attractive forces between connected nodes, and a weakly repulsive force between all nodes. The result is that facts which co-occur in many of the same pages are placed close to each other, and close to the pages that mention them. The analysis revealed that there are two subgroups of pages (the top and bottom), which contain a concentration of different sets of facts, while another group of pages in the middle that contain both groups of facts. Furthermore, the visualization reveals common and rare facts, and how they co-occur across the pages. Such visualizations (and related quantitative network measures used to verify the visual observations) therefore help to reveal new regularities about information scatter which are often concealed in aggregate measures such as frequency distributions.

Explanations for Information Scatter

While there are many studies that quantify information scatter, there are relatively few explanations for

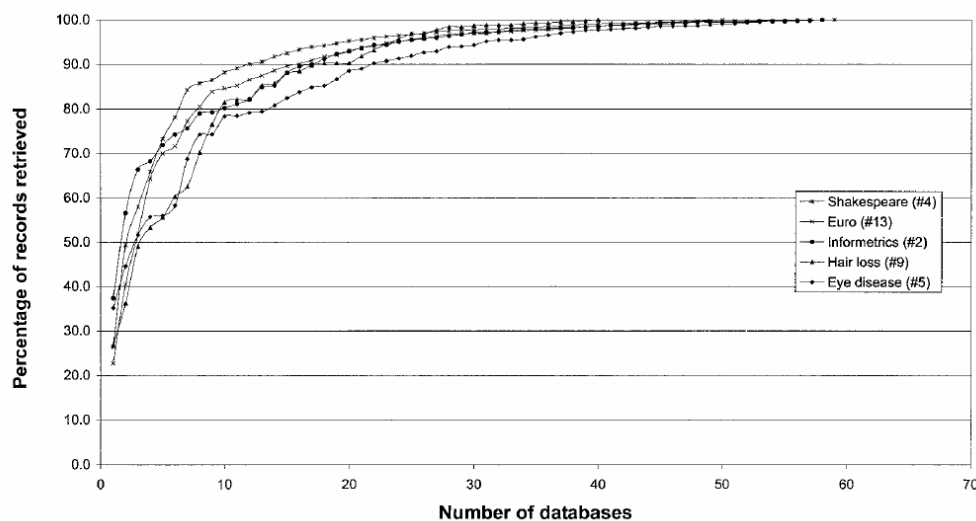


Figure 2. Distribution of the percentage of journal articles retrieved as a function of the number of databases accessed (Hood and Wilson, 2001). Reprinted with permission of John Wiley & Sons, Inc.

why information scatter occurs. The scatter of articles across journals has been explained by the fact that each field has a core set of journals where authors publish. However, because topics often overlap with other fields (interdisciplinarity or multidisciplinary), authors from a field may publish in journals that are related as well as distant to their main field. In that sense, the core journals of one field becomes the peripheral journals of another field, resulting in the universality of the Bradford law of scattering [18]. At a different level of granularity, Hood and Wilson [13] proposed that information scatter could be related to the interdisciplinarity of the topic. The more interdisciplinary a topic, the higher probability that the authors could choose to publish results of the same topic in journals of different fields, resulting in the high scatter of that topic across journals and databases.

Bhavnani [15] found that facts (e.g., High UV exposure increases your risk of getting melanoma) about five healthcare topics (e.g., melanoma risk and prevention) were scattered across high-quality websites specializing in that topic. There were many pages which had few facts, few pages that had many facts, and no pages that had all the facts. Furthermore, the analysis revealed that underlying this distribution were three different page profiles that varied in fact density and role: *General pages* which contain many facts in medium amount of detail and play the role of providing an overview of the topic (e.g., What Are the Risk Factors for Melanoma?); *Specific pages* which contain few facts in a large amount of detail and play the role of providing detailed description of a few facts (e.g, Sunscreens and Prevention of Malignant Melanoma); and *Sparse pages* which contain few facts in a small amount of detail and were pages that were of broader topics (e.g., Skin Cancer) with a brief reference to the search topic.

The above observations led to the *information saturation* model [19] to explain this phenomenon. In this hypothetical model, webpage authors follow a process of *accumulation* to progressively add facts in detail to a page, until a length and detail saturation threshold is reached. At such a threshold, because the webpage becomes unwieldy and difficult to read, webpage authors heed design guidelines by removing detail about facts from these pages through the process of *abstraction* resulting in general pages. Concurrently, they also might be creating new pages to elaborate particular facts in high detail through the process of *specialization* resulting in specific pages. Finally as a topic becomes more important, the

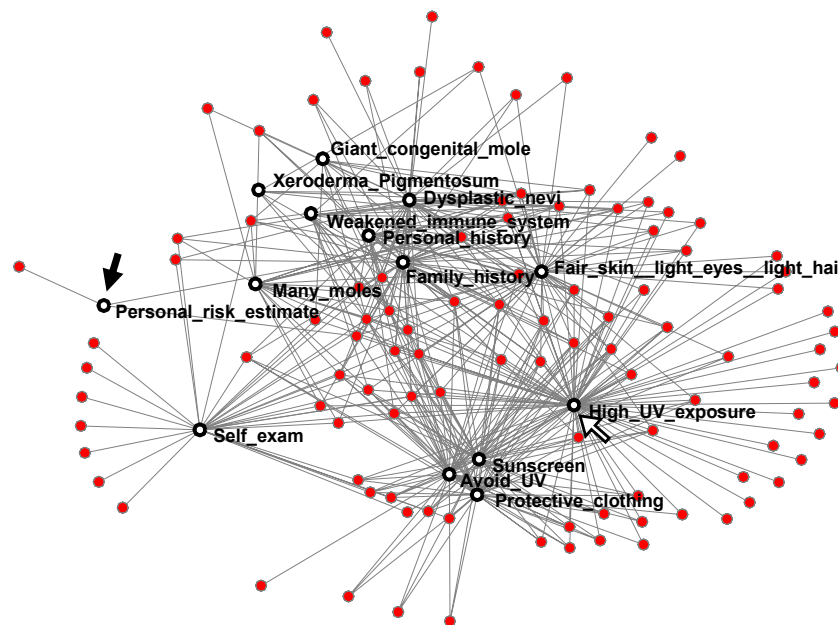


Figure 3. The scatter network for risk/prevention showing how 14 facts (white labeled nodes) occur in 108 relevant pages (solid nodes). The white arrow points to a common fact, and the black arrow points to a rare fact (Adamic et al., 2007). Reprinted with permission of IOP Publishing Ltd.

topic *permeates* otherwise irrelevant pages resulting in sparse pages. The above processes could lead to the creation of a large number of specific pages and sparse pages, while constraining the total number of general pages. However, while there have been quantitative models that generate different informetric phenomena such as the distribution of papers per author [20], there appears to be no research that has quantitatively modeled phenomena related to information scatter.

Implications of Information Scatter for Search and Design

While there have been claims that Bradford's law has been useful for selecting journals for libraries or databases, there are actually no references about how it has been applied in practical library and information sciences [9]. Some researchers have explored the implications for the phenomenon of information scatter to search strategies, and the design of systems to help users find comprehensive information. Bates [21] proposed that because the core, middle, and peripheral regions of the Bradford distribution have different densities of relevant articles, they imply the use of different search strategies. For example, when searching for articles, users should identify the core set of journals for that topic. Because there is a high density of relevant articles in the core, browsing could be sufficient to find many relevant articles. In contrast, because the middle region has a lower density of relevant articles, users need to use queries that exploit systematic organizations of articles such as indexes. Finally, in the peripheral region where articles are very scattered, users need to use strategies such as *citation following* to quickly find the relevant articles.

While the above search strategies are implied by the density of articles *across* the different regions of the Bradford distribution, Bhavnani [15] leveraged the idea of information density *within* webpages through the *general-specific-sparse* search strategy. This strategy recommends that users first find and read general pages about a topic to get an overview of the topic, and to ensure that they do not miss any facts. Next (as often prescribed by earlier authors such as Kirk [22]), users should find specific pages that specialize in the facts they found within the general pages, to help elaborate facts of interest. Finally, they should broaden the search by finding and reading sparse pages to understand how the topic of interest relates to other topics.

The above strategy was operationalized in a website called the *Strategy Hub* [23] to help users finding comprehensive healthcare information. In this system webpages were organized in terms of general, specific, and sparse pages, and users were guided to read pages that followed that strategy. In a controlled experiment, users of the Strategy Hub were found to find more comprehensive information in the same amount of time compared to equivalent users of MedlinePlus and Google.

Future Research Challenges

Future research challenges include (1) building models to understand the process of information scatter, and (2) algorithms to help users gather highly scattered information. As described earlier, although a few researchers have proposed explanations for information scatter, none of these have been formalized into a model and tested to demonstrate the process of information scatter. Future research should therefore develop computational models to simulate information scatter over time so that the phenomenon can be precisely understood. The second area of potential future research is to leverage a precise understanding of information scatter to develop new algorithms to either automatically aggregate different types of scattered information, or new algorithms that work interactively with users to help them find comprehensive information.

Additional Readings

Bar-Ilan, J. (2008). Informetrics. In: *Encyclopedia of Library and Information Sciences*. Marcia Bates, Mary Niles Maack and Miriam Drake (editors), 2nd edition. xxxx

Rouseau, R. (2008). Informetric laws. In: *Encyclopedia of Library and Information Sciences*. Marcia Bates, Mary Niles Maack and Miriam Drake (editors), 2nd edition. xxxx

Acknowledgements

This research was supported by NIH grant # UL1RR024986.

References

- (1) Bradford, S. C. Sources of Information on Specific Subjects. *Engineering* **1934**, 137 (3550), 85-86. [Reprinted in: Bradford, S.C. (1948). *Documentation*. London: Crosby Lockwood.]
- (2) Campbell, F. *The theory of national and international bibliography*. London: Library Bureau, 1896.
- (3) Wilson, C.S. Informetrics. *Annual Review of Information Science and Technology* **1999**, 34, 107-247.
- (4) Vickery, B. C. Bradford's Law of Scattering. *Journal of Documentation* **1948**, 4(3), 198-203.
- (5) Chen, Y. and Leimkuhler, F.F. A relationship between Lotka's Law, Bradford's Law and Zipf's Law. *Journal of the American Society for Information Science* **1986**, 37(5), 307-314.
- (6) Zipf, G.K. *Human behavior and the principle of least-effort; an introduction to human ecology*. Cambridge, MA: Addison-Wesley, 1949.
- (7) Lotka, A.J. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* **1926**, 16 (12), 317-323.
- (8) Bar-Ilan, J. Informetrics at the beginning of the 21st century--A review. *Journal of Informetrics* **2008**, 2(1), 1-52.
- (9) Nicolaisen, J. & Hjørland, B. Practical potentials of Bradford's law: A critical examination of the received view. *Journal of Documentation* **2007**, 63(3), 359-377.
- (10) Wilson, C.S. The formation of subject literature collections for bibliometric analysis: The case of the topic of Bradford's Law of Scattering. Sydney, Australia: The University of New South Wales, 1995. PhD dissertation.
[<http://unsworks.unsw.edu.au/vital/access/manager/Repository/unsworks:412>]
- (11) Tenopir, C. Evaluation of database coverage: A comparison of two methodologies. *Online Review* 1982, 6, 423-441.
- (12) Lancaster, F. W. and Lee, J-L. Bibliometric techniques applied to issue management: A case study. *Journal of the American Society for Information Science* **1985**, 36(6), 389-397.
- (13) Hood, W., & Wilson, C.S. The scatter of documents over databases in different subject domains: How many databases are needed? *Journal of the American Society for Information Science* **2001** 52(14), 1242-1254.
- (14) Bhavnani, S.K. The Retrieval of Highly Scattered Facts and Architectural Images: Strategies for Search and Design. *Automation in Construction* **2005**, 14(6), 724-735.

- (15) Bhavnani, S.K. Why is it difficult to find comprehensive information? Implications of information scatter for search and design. *Journal of the American Society for Information Science and Technology* **2005**, 56(9), 989-1003.
- (16) Over, P. TREC-6 Interactive track report. In NIST Special Publication, The Seventh Text Retrieval Conference; E. M. Voorhees, 1., D. K. Harman, 2. Eds.; **1998**. 500-242.
- (17) Adamic, L.A., Bhavnani, S.K., and Xiaolin, S. Scatter networks: A new approach for analyzing information scatter on the web. *New Journal of Physics (Special Issue on Complex Systems)* **2007**, 9, 231.
- (18) Garfield, E. Bradford's Law and Related Statistical Patterns. *Essays of an Information Scientist* **1980**, 4, 476-483.
- (19) Bhavnani, Suresh K. and Peck, Frederick A. Towards a Model of Information Scatter: Implications for Search and Design. In Grove, Andrew, Eds. *Proceedings 69th Annual Meeting of the American Society for Information Science and Technology (ASIST)* **2006**, 43, Austin (US).
- (20) Goldstein, M. L., Morris, S. A., & Yen, G. G. Group-based Yule model for bipartite author-paper networks. *Physical Review* **2005**, E, 71.
- (21) Bates, M. J. Speculations on browsing, directed searching, and linking in relation to the Bradford Distribution. In: H. Bruce, R. Fidel, R. Ingwersen, and P. Vakkari (Eds) *Emerging Frameworks and Methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)*. Greenwood Village, CO: Libraries Unlimited, 2002; 137-150.
- (22) Kirk, T. Problems in library instruction in four-year colleges. In: Lubans, John, Jr. *Educating the library user*; 1st. ed. New York, NY: R.R. Bowker, 1974; 83-103.
- (23) Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., and Strecher, V.J. Strategy hubs: Domain portals to help find comprehensive information. *Journal of the American Society for Information Science and Technology* **2006**, 57(1), 4-24.