Applying Bradford's Law of Scattering in Digital Libraries

Philipp Mayr
GESIS / Social Science Information Centre, Bonn (Germany)

Doctoral candidate:

Philipp Mayr

Social Science Information Centre
Informationszentrum Sozialwissenschaften
Department: Research and Development
Lennéstr. 30, D - 53113 Bonn (Germany)
http://www.gesis.org/en/iz/index.htm

Telefone: +49-228-2281-181
Fax: +49-228-2281-121
Email: mayr@iz-soz.de
http://www.ib.hu-berlin.de/~mayr/

Supervisors:

Prof. Dr. Walther Umstätter

Humboldt University Berlin
Unter den Linden 6
Institut für Bibliotheks- und Informationswissenschaft
Dorotheenstraße 26, D-10117 Berlin (Germany)
Email: h0228kdm@rz.hu-berlin.de
http://www.ib.hu-berlin.de/%7Ewumsta/infopub/

Prof. Dr. Jürgen Krause

Scientific director Social Science Information Centre
Professor for Computer science University Koblenz-Landau
Lennéstr. 30, 53113 Bonn (Germany)
Tel.: +49 -228-2281-145
Email: jk@bonn.iz-soz.de
http://www.gesis.org/IZ/Krause/

Overview

The purpose of this project is the application of the bibliometric Bradford Law of Scattering (BLS) for generating core document sets for subject specific questions. BLS is used to re-order result sets and discover interdisciplinary properties of result sets from distributed searches.

Introduction

The background of the research project is that distributed search across multiple databases over the WWW will automatically generate large and heterogeneous document sets for subject specific questions. As a result, users have to deal with a huge amount of documents from different domain, and also for specific research topics. The perceived expectations of users searching the web are that system architects should list the most relevant or important documents in the result list first and additionally create flexible environments especially for scientific retrieval. More and more approaches appear which draw on advanced methods to produce quality results. Google PageRank and Google Scholar's citation count are only two famous examples for informetric-based mechanisms applied in Internet search engines or digital libraries to satisfy user demands.

Figure 1 shows the classical two levels of treatment of vagueness in information retrieval. Vagueness 1 (V1) – the mapping between user questions and document terms – will not be touched in our project (see (Petras, 2006) for a V1-treatment experiment). Our project is concentrated on treating the vagueness/heterogeneity between different document collections (mainly bibliographic databases and library catalogues) and their controlled vocabularies on a semantic level (see V2/V3 in Figure 1) which arises in distributed search.

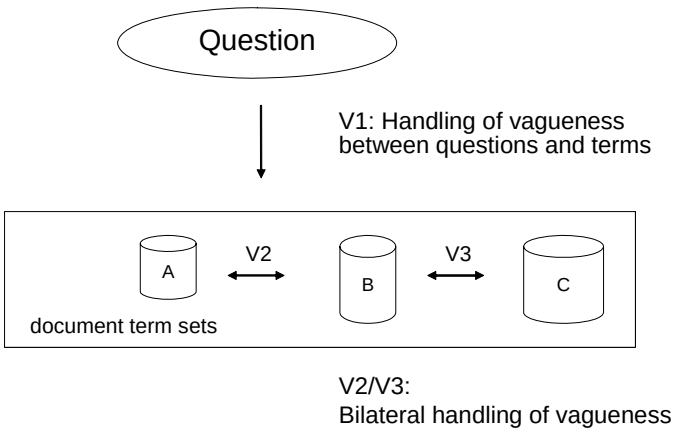

Figure 1: Treatment of vagueness in information retrieval (two-step methodology)

Objectives of our project[1] are: 1) to explore suitable methods in model building in an environment of scientific documents (e.g. the German science portal vascoda[2]) which mostly are indexed with various controlled vocabularies (thesauri, subject headings, etc) and 2) to create and apply modules that treat semantic heterogeneity (V2/V3) between bibliographic collections. Cross-concordances[3] are one way to treat semantic heterogeneity (Krause, 2003) in searching heterogeneous indexed collections via one search vocabulary (Hellweg et al., 2001, Zeng and Chan, 2004). Our project will provide a web service which transforms

[1] http://www.gesis.org/en/research/information_technology/komohe.htm

[2] http://www.vascoda.de/

[3] Cross-concordances are bilateral mappings of controlled descriptors with the objective to search a semantic concept in all semantic integrated collections.

descriptor terms of known vascoda vocabularies into other vocabularies using cross-concordances.

One direction in the project will be the application of the bibliometric Bradford Law of Scattering (BLS) for generating core document sets for subject specific questions BLS is used to re-order result sets and discover interdisciplinary properties of result sets from distributed searches (Tenopir, 1982).
An extensive review of the literature of BLS (Bradford, 1948, Bradford, 1934) is provided by (Lockett, 1989). BLS is a common known bibliometric law which has drawn a lot of attention in information and library science research (Vickery, 1948, Brookes, 1969, Worthen, 1975, Buckland, 1972, Brookes, 1968, Vickery, 1948). BLS is still under discussion as shown by recent papers (Nicolaisen and Hjørland, 2007, Mayr and Umstätter, 2007, Umstätter, 2005, Hjørland and Nicolaisen, 2005, Bates, 2002, Hood and Wilson, 2001).

Research questions
The application of BLS in our project has two different perspectives:

1) BLS as a supporting mechanism for information retrieval

The paper from Bates (2002) is interesting in this context because it brings together BLS and information seeking behavior.

„… the key point is that the distribution tells us that information is neither randomly scattered, nor handily concentrated in a single location. Instead, information scatters in a characteristic pattern, a pattern that should have obvious implications for how that information can most successfully and efficiently be sought." (Bates, 2002)

Bates applies conceptually different search techniques (directed searching, browsing and linking) to the Bradford zones. Bates postulates the Bradford nucleus for browsing, the following zone for directed searching with search terms and further zones for linking.

We are focusing on an automatic change between directed searching (enhanced by treatment of semantic heterogeneity) into browsing. Starting with a subject specific descriptor search (see step 1 identify in Figure 2), we will connect the query with our heterogeneity modules to transfer descriptor terms into a multi-database scenario. In the second step, the result lists from the different databases will be combined and sorted according to Bradford's method (most productive journals for a topic first). After this step we have a bradfordized list of journal articles (White, 1981). Step 3 is the extraction of a result set of all documents in the Bradford nucleus which can be delivered for browsing. This automatically generated browsing modus can be compared to Bates search technique "journal run".

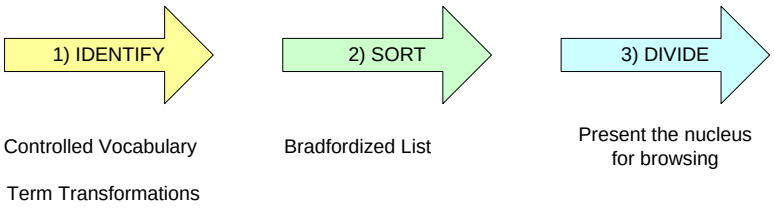

Figure 2: Application of Bradfordizing

2) BLS as a way to analyze the consequence of using modules of semantic treatment of heterogeneity

The second way of applying BLS is more informetric. We postulate that using semantic treatment of heterogeneity will enlarge and complete an interdisciplinary document space which is per se only a plus in document recall. We further believe that BLS can help to analyze and evaluate the effect of automatically transferring controlled terms while subject

searching. A metric approach based on BLS can hopefully be used to describe scattering, interdisciplinarity and other term transformation effects.

General questions:

- Is a re-ranking of documents according to the Bradford zones an added value for users? The reduction of content to the most pertinent sources (extracting the nucleus) can be a helpful access mechanism for browsing. The evaluation of the utility of such a mechanism is still a desideratum.
- Are the documents in the nucleus of a bradfordized list (core journals) more relevant for a topic than results in further zones? A study by (Pontigo and Lancaster, 1986) concluded that less productive journals are not necessarily of lower quality but mostly less cited. This has to be proven by intellectual assessments (analogue to the TREC or CLEF studies) of different user groups (e.g. experts, novice searchers, information scientists).
- Can BLS be applied to other document sources than journal articles? A paper by Worthen (1975) and our own analyses (see Figure 3) show that monograph literature can be successful bradfordized. Other document types (proceedings, grey literature etc.) have to be equally proven.
- Can BLS be applied for any term-based queries e.g. searching an author name or an uncontrolled search term.
- Can BLS be found in all subject domains of our scenario? There are lots of examples of applications of BLS in various disciplines, natural sciences and social sciences (Peritz, 1990).
- Can the typical Groos droop in Bradford distributions be attenuated when searching with term transformations?

Methodology

We focus on a mix of methodologies.

- Bradfordizing (White, 1981) as a sorting mechanism for connected databases in our distributed scenario
- Empirical analysis of the results for subject specific topics and questions
- Assessments of document relevance (comparing a subject search with and without search term transformation) for various user types
- User tests of a prototypical implementation of the bradfordized browing modus

Motivation

The main motivation for this application is to get feedback from professional information science researchers and discuss the applicability and utility of the approach in this informetric forum. I would like to learn from the experiences of the participants while using BLS or other statistical laws in informetric analyses or informetric-enhanced information systems.

Concrete questions to the forum:

- Are there any experiences or empirical studies with users using Bradfordizing?
- How could a metric be constituted for measuring interdisciplinarity in distributed bibliographic document sets?
- Are there other experiences in applying BLS to document types other than journal articles?

Figure 3 shows an analysis of documents following BLS. All analyzed document collections and document types (journal articles and monographs) show a similar distribution.

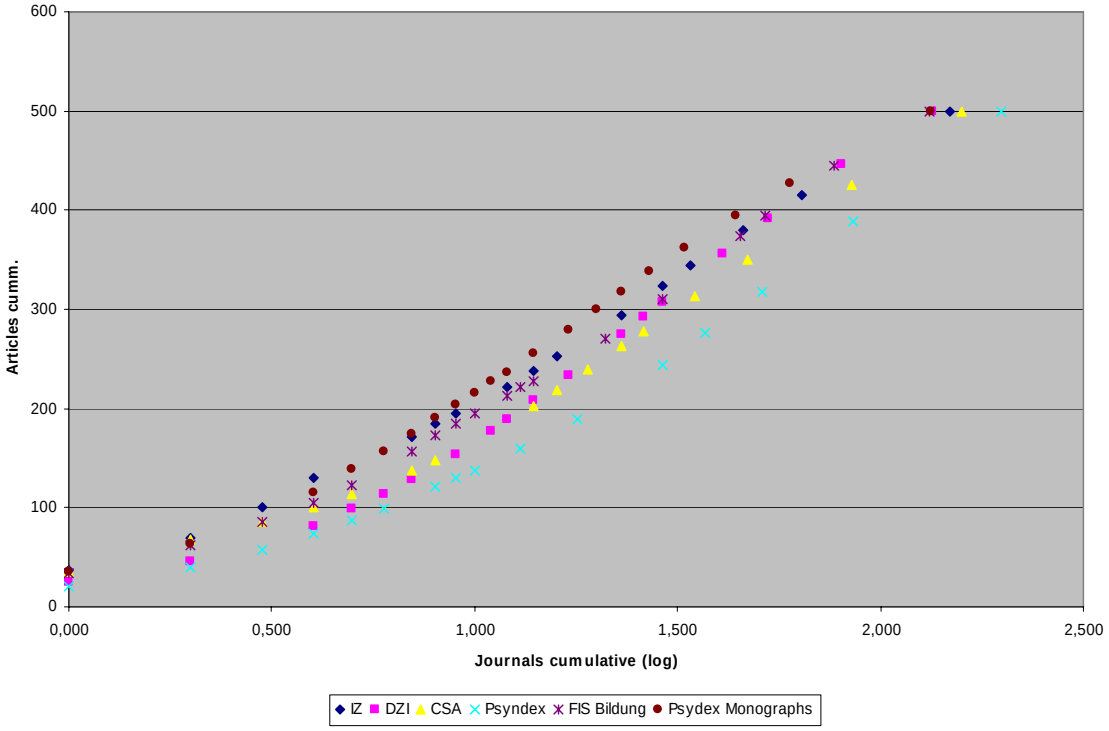

Figure 3: Bradford graphs for different bibliographic databases[4] (the topic "youth research" for various journals articles distributions and a distribution of monographs)

References

BATES, M. J. (2002) Speculations on Browsing, Directed Searching, and Linking in Relation to the Bradford Distribution. IN BRUCE, H., FIDEL, R., INGWERSEN, P. & VAKKARI, P. (Eds.) *Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4).*

BRADFORD, S. C. (1934) Sources of information on specific subjects. *Engineering,* 137**,** 85-86.

BRADFORD, S. C. (1948) *Documentation,* London, Lockwood.

BROOKES, B. C. (1968) The derivation and application of the Bradford-Zipf distribution. *Journal of Documentation,* 24**,** 247-265.

BROOKES, B. C. (1969) Bradford's Law and the Bibliography of Science. *Nature,* 224**,** 953-956.

BUCKLAND, M. K. (1972) Are Obsolescence and Scattering related? *Journal of Documentation,* 28**,** 242-246.

HELLWEG, H., KRAUSE, J., MANDL, T., MARX, J., MÜLLER, M. N. O., MUTSCHKE, P. & STRÖTGEN, R. (2001) *Treatment of Semantic Heterogeneity in Information Retrieval,* Bonn, IZ Sozialwissenschaften.

HJØRLAND, B. & NICOLAISEN, J. (2005) Bradford's law of scattering: ambiguities in the concept of "subject". IN CRESTANI, F. & RUTHVEN, I. (Eds.) *5th International Conference on Conceptions of Library and Information Science.* Springer-Verlag.

HOOD, W. W. & WILSON, C. S. (2001) The scatter of documents over databases in different subject domains: how many databases are needed? *Journal of the American Society for Information Science and Technology,* 52**,** 1242-1254.

KRAUSE, J. (2003) Standardization, heterogeneity and the quality of content analysis: a key conflict of digital libraries and its solution. *IFLA 2003, World Library and Information Congress: 69th IFLA General Conference and Council.* Berlin.

[4] IZ = a social science database at the institute, DZI = an other social science database at the institute, CSA = social science databases from Cambridge Scientific Abstracts; Psyndex = psychological database, FIS Bildung = pedagogic database; Psyndex Monographs = psychological database restricted on the document type monographs.

LOCKETT, M. W. (1989) The Bradford distribution. A review of the literature, 1934-1987. *Library and Information Science Research,* 11**,** 21-36.
MAYR, P. & UMSTÄTTER, W. (2007) Why is a new Journal of Informetrics needed? *Cybermetrics,* 11.
NICOLAISEN, J. & HJØRLAND, B. (2007) Practical potentials of Bradford's law: A critical examination of the received view. *Journal of Documentation,* 63.
PERITZ, B. C. (1990) A Bradford distribution for bibliometrics. *Scientometrics,* 18**,** 323-329.
PETRAS, V. (2006) Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages. Berkeley, USA, University of California, Berkeley.
PONTIGO, J. & LANCASTER, F. W. (1986) Qualitative aspects of the Bradford distribution. *Scientometrics,* 9**,** 59-70.
TENOPIR, C. (1982) Distributions of citations in databases in a multidisciplinary field. *Online Review,* 6**,** 399-419.
UMSTÄTTER, W. (2005) Anmerkungen zu Birger Hjørland und Jeppe Nicolaisen: Bradford's Law of Scattering: Ambiguities in the Concept of "Subject". *Libreas*.
VICKERY, B. C. (1948) Bradford's law of scattering. *Journal of Documentation,* 4**,** 198-203.
WHITE, H. D. (1981) 'Bradfordizing' search output: how it would help online users. *Online Review,* 5**,** 47-54.
WORTHEN, D. B. (1975) The application of Bradford's law to monographs. *Journal of Documentation,* 31**,** 19-25.
ZENG, M. L. & CHAN, L. M. (2004) Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. *Journal of the American Society for Information Science and Technology,* 55**,** 377-395.