Bradfordizing effects

Philipp Mayr

GESIS - Leibniz Institute for the Social Sciences
Lennéstr. 30, 53113 Bonn, Germany

E-Mail: philipp.mayr@gesis.org

Abstract

The purpose of this paper[1] is to apply and evaluate the bibliometric method Bradfordizing for information retrieval (IR) experiments. Bradfordizing is used for generating core document sets for subject-specific questions and to re-order result sets. The method will be applied and tested in a controlled scenario of scientific literature databases from social and political sciences, economics, psychology and medical science and 164 standardized IR topics. The IR tests show that relevance distributions after re-ranking improve at a significant level if articles in the core are compared with articles in the succeeding zones.

1 Introduction

Distributed search across multiple databases will automatically generate large and heterogeneous document sets for subject-specific questions with the result that users are confronted with a massive load of documents from different scientific domains, even for specific research topics. The perceived expectations of users searching the web are that system architects should list the most relevant or valuable documents in the result list first (so-called relevance ranking). More approaches appear that draw on advanced methods to produce qualitative results and alternative views on document spaces. Google PageRank and Google Scholar's citation count are just two popular examples for

[1] This paper is a short version of the paper presented at COLLNET meeting 2008 in Berlin (see also Mayr, submitted).

informetric-based mechanisms applied in Internet search engines. Similar techniques can and should be applied in digital libraries (DL) to satisfy user demands. The KoMoHe[2] project at GESIS was the starting point and background of the following approach. This paper should be seen as an argument and example for alternative (informetric) re-ranking methods applied in text-based retrieval systems or DL (see Mayr et al., 2008 for an outline).

Bradford's Law of Scattering (BLS), a well known bibliometric law, has garnered a lot of attention in information and library science research. Numerous examples of applications of Bradford law can be found among various disciplines including natural and social sciences. This empirical law appears to be a very robust and commonly appearing phenomenon in most of the current literature databases and bibliographies. Bates' paper (2002) is interesting specifically in our context because it brings together BLS, information seeking behavior and IR. "… *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 applies the search techniques of directed searching, browsing and linking to the classical three Bradford zones. Whereby, she postulates utilization of the Bradford nucleus (core) for browsing, the following zone (z2) for directed searching with search terms and further zones (z3) for linking.

Our goal is to go directly and automatically from directed searching into browsing. Starting with a subject-specific descriptor search, we will treat the query with our heterogeneity modules (Mayr & Petras, 2008) to transfer descriptor terms into a multi-database scenario. In a second step, the result lists from the different databases are 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. Step 3 is the extraction of a result set of all documents in the Bradford nucleus which can be delivered for browsing. This

[2] "Competence Center Modeling and Treatment of Semantic Heterogeneity", see http://www.gesis.org/forschung-lehre/programme-projekte/informationswissenschaften/projektuebersicht/komohe/

browsing modus, based on automatically bradfordized lists, can be compared to the search technique Bates terms “journal run.”

2 Research questions

We seek to answer the following research questions:

1) Is a re-ranking of documents according to the Bradford law (journal productivity) an added value for users? The re-ranking of content to the most frequent sources (extracting the nucleus) can, for example, be a helpful access mechanism for browsing and initial search stages. Evaluation of the utility of such a mechanism is still a desideratum.

2) Are the documents in the nucleus of a bradfordized list (core journals show high productivity for a topic) more relevant for a topic than items in succeeding zones with lower productivity? This requires proving on a larger scale via intellectual assessments by different user groups (e.g. experts, novice searchers, information scientists).

3) Can Bradfordizing be applied to document sources other than journal articles? Few analyses show that monograph literature can be successfully bradfordized. But is this a utility? Other document types (proceedings, grey literature etc.) have to be equally proven.

4) Can Bradfordizing be used to create an alternative view on search results? Compared to traditional text-oriented ranking mechanisms, our informetric re-ranking method offers a completely new view on results sets, which have not been implemented and tested in heterogeneous database scenarios with multiple collections to date.

3 Methods

We focus on a mix of methodologies:

- Bradfordizing as a sorting mechanism for documents in our distributed database scenario. White explains the Bradfordizing procedure: “… *That is*

sorting hits (1) by the journal in which they appear, and then sorting these journals not alphabetically by title but (2) numerically, high to low, by number of hits each journal contains. In effect, this two-step sorting ranks the search output in the classic Bradford manner, so that the most productive, in terms of its yield of hits, is placed first; the second-most productive journal is second; and so on, down through the last rank of journals yielding only one hit apiece." (1981: p. 47). Our study analyzed scientific literature from social and political sciences, economics, psychology and medical science databases (SOLIS, SoLit, USB Köln Opac, CSA Sociological Abstracts, World Affairs Online, Psyndex and Medline) in exactly this way: 1) Searching documents, 2) sorting or re-ranking documents via Bradfordizing, 3) dividing documents into three, equally-sized zones.

- Intellectual assessments of document relevance were performed following the classical IR evaluation experiments at TREC and Cross-Language Evaluation Forum (CLEF). That followed an empirical analysis of the retrieval results for subject-specific topics and questions. We retrieved, analyzed and intellectually assessed 164 different standardized topics which yielded more than 96,000 documents from all above domains. More than 51,000 assessed documents could be bradfordized.

- The utility of the nucleus/core was investigated in a simple user-test based on 24 qualitative interviews.

4 Results[3]

An evaluation of the method and its effects was carried out in two laboratory-based information retrieval experiments (CLEF and KoMoHe) using a controlled document corpus and human relevance assessments. The results show that Bradfordizing is a very robust method for re-ranking the main document types (journal articles and monographs) in today's digital libraries (DL). The IR tests show that relevance distributions after re-ranking improve at a significant

[3] Detailed results of the research can be found in Mayr (submitted).

level if articles in the core are compared with articles in the succeeding zones. The items in the core are significantly more often assessed as relevant, than are items in zone 2 or zone 3. The largest increase in precision can typically be observed between core and zone 3. Bradfordizing can successfully be applied in a set of scientific literature databases and holds true in different domains and document types. A value-added for this re-ranking method can be empirically demonstrated in terms of precision improvements on a significant level. Users are intuitively satisfied with the re-ranked results. The results can also be seen as a concretion of Bradford Law in so far as Bradford did not postulate or observe a relevance advantage in the core. The results show that articles in core journals are valued more often relevant then articles in succeeding zones. This is an extension to the original conception of relevance distribution in the zones by Bradford. The relevance advantage in the core can probably be explained in that a) core journals publish more state-of-the-art articles, b) core journals are more often peer-reviewed and c) core journals cover more aspects of the searched topic than journals in the peripheral zones.

5 Further research

Further research will focus on the implementation and evaluation of the method in a live system with different modules for improving retrieval (see Mayr et al., 2008). The exploration of the side effects and bias of this promising re-ranking method will be a next step. The application of other evaluation methods (e.g. evaluation of full texts instead of metadata) would be highly desired.

6 References

Bates, Marcia J. (2002): Speculations on Browsing, Directed Searching, and Linking in Relation to the Bradford Distribution.

Mayr, P. (submitted): Re-Ranking auf Basis von Bradfordizing für die verteilte Suche in Digitalen Bibliotheken. Dissertation. Philosophische Fakultät I, Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin. 238 pages.

Mayr, P. (2008). An evaluation of Bradfordizing effects, Proceedings of WIS 2008, Berlin, Fourth International Conference on Webometrics, Informetrics and

Scientometrics & Ninth COLLNET Meeting, Humboldt-Universität zu Berlin. URL: http://www.collnet.de/Berlin-2008/MayrWIS2008ebe.pdf

Mayr, P.; Mutschke, P.; Petras, V. (2008). Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking. In: Library Review 57, No. 3, pp. 213-224.

Mayr, P.; Petras, V. (2008). Cross-concordances: terminology mapping and its effectiveness for information retrieval. In: IFLA World Library and Information Congress. Québec, Canada URL: http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf

White, H. D. (1981). 'Bradfordizing' search output: how it would help online users. In: Online Review 5, No. 1, pp. 47-54.