

Bradfordizing mit Katalogdaten

Philipp Mayr

GESIS – Leibniz-Institut für Sozialwissenschaften

Nutzer erwarten für Literaturrecherchen in wissenschaftlichen Suchsystemen (u. a. OPACs) einen möglichst hohen Anteil an relevanten und qualitativen Dokumenten in den Trefferergebnissen. Insbesondere die Reihenfolge und Struktur der gelisteten Ergebnisse (Ranking) spielt, neben dem direkten Volltextzugriff auf die Dokumente, für viele Nutzer inzwischen eine entscheidende Rolle. Abgegrenzt wird Ranking oder Relevance Ranking von sog. Sortierungen z. B. nach dem Erscheinungsjahr der Publikation, obwohl hier die Grenze zu „nach inhaltlicher Relevanz“ gerankten Listen konzeptuell nicht sauber zu ziehen ist. Das Ranking von Dokumenten führt letztlich dazu, dass sich die Benutzer fokussiert mit den oberen Treffermengen eines Suchergebnisses beschäftigen. Der mittlere und untere Bereich eines Suchergebnisses wird häufig nicht mehr in Betracht bezogen. Auf Grund der Vielzahl an relevanten und verfügbaren Informationsquellen ist es daher notwendig, Kernbereiche in den Suchräumen zu identifizieren und diese anschließend dem Nutzer hervorgehoben zu präsentieren.

Die qualitativen Verfahren der traditionellen Fachinformationsanbieter (z. B. Bibliotheken) zeigen bekanntlich bei den Punkten Ranking und Volltextzugriff Schwächen, überzeugen aber vor allem durch ihre Stringenz, in diesem Fall die selektive Aufnahme von qualitätsgeprüften Dokumenten, Erstellung von qualitativen Metadaten (z.B. Autoren- und Titelansetzungen) sowie die inhaltliche Erschließung der Dokumente mit kontrollierten Vokabularen. Aufgrund der Beschaffenheit der Dokumente in klassischen Informationssystemen wie z. B. bibliographischen Fachdatenbanken oder auch OPACs greifen die Ranking-Verfahren des Web Information Retrieval (sog. Suchmaschinentechologie) genauso wenig wie die herkömmlichen textstatistischen Information Retrieval-Verfahren (z.B. tf-idf). Alternative, nicht-textorientierte Ansätze, wie z. B. der Einsatz von Autorenzentralitätsmaßen beim Ranking (Mutschke, 2004), aber auch das in diesem Artikel vorgestellte Verfahren, zeigen, dass die spezifische Struktur der Literaturnachweise und die Qualität der Metadaten für alternative Mehrwertdienste gewinnbringend eingesetzt werden kann (siehe Mayr et al. 2008).

Das in diesem Artikel vorgestellte Verfahren Bradfordizing (White, 1981) setzt auf die bibliometrische Regelmäßigkeit des Bradford Law of Scattering (Bradford, 1948) und nutzt diese Regelmäßigkeit zum Ranking bzw. ReRanking von Dokumenten (Mayr, 2009). Dem Bradford Law of Scattering (BLS) liegt zugrunde, dass sich die Literatur zu einem beliebigen Fachgebiet bzw. -thema z. B. in einer Bibliographie, in Bereiche mit unterschiedlichen Dokumentenkonzentrationen unterteilen lässt (siehe Abb. 1). So besteht zwischen den Zeitschriften eines Forschungsthemas und den Artikeln in diesen Zeitschriften eine quantifizierbare Relation, die Bradford als erstes beschrieben hat. Dem Kernbereich mit hoher Konzentration der Literatur folgen Bereiche mit zunächst mittlerer und geringer Konzentration, die jeweils die gleiche Menge an Zeitschriftenartikeln beinhalten wie der Kernbereich. Im Prinzip stellt das BLS lediglich eine Präzisierung der 80:20-Regel für den Bereich der Zeitschriftenliteratur dar. 80 % der Nachfrage an Literatur können mit etwa 20 % des Bestandes in einer Bibliothek abgedeckt werden (vgl. Umstätter, 2005). Anders formuliert: 80 % der Zeitschriftenartikel zu einem

Thema finden sich in 20 % der Zeitschriften, die zu diesem Thema publizieren. BLS hat damit Parallelen zu Paretos Beobachtungen auf dem Gebiet der Einkommensverteilungen.

Das folgende idealisierte Beispiel in Abbildung 1 visualisiert die drei Zonen nach einer Bradfordizing-Analyse. Die Dokumentmenge umfasst in diesem Beispiel 450 Zeitschriftenartikel zu einem Forschungsthema, die in dem Fall auf insgesamt 39 unterschiedliche Zeitschriften verteilt sind (150 Artikel in jeder der drei Zonen). Das Kumulieren der Artikelzahlen nach dem Bradfordizing hat in diesem Fall ergeben, dass die ersten drei Zeitschriften zusammen 150 Artikel ergeben, also das erste Drittel der Gesamtdokumentenzahl von 450 Dokumenten. Diese ersten 150 Dokumente in den drei Zeitschriften definieren damit den Nukleus oder Core für dieses Topic. Die drei Zeitschriften werden „Core Journals“ oder Kernzeitschriften genannt. Für das zweite Drittel bzw. Zone 2 werden 9 (3*3 Zeitschriften) weitere Zeitschriften und für das dritte Drittel bzw. Zone 3 werden 27 (3*3*3 Zeitschriften) Zeitschriften benötigt, um die Menge von jeweils 150 Zeitschriftenartikeln zu erreichen.

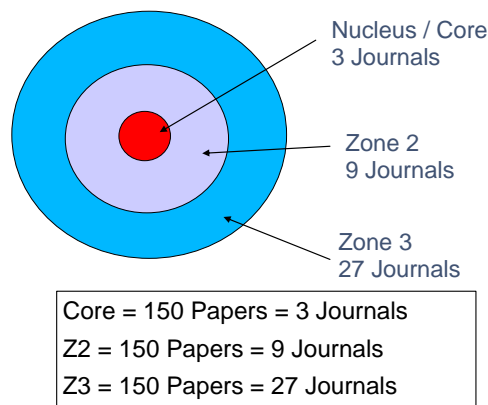


Abb. 1: Ergebnis eines idealisierten Dokumentenpools für eine Fragestellung nach Bradfordizing. Einteilung in drei Bradford-Zonen (Core, Zone 2 und Zone 3).

Die Fokussierung auf Bradfordizing erscheint vielversprechend, weil dieses Verfahren per se universell und disziplinübergreifend angelegt ist und zuverlässig sowohl innerhalb einer Datenbank, einer Domäne (mehreren Datenbanken zu einem Fachgebiet) als auch zwischen Domänen beobachtet werden kann. Bradfordizing lässt sich u.a. erfolgreich auf Monographien z.B. in OPACs oder Fachdatenbanken anwenden (vgl. Worthen, 1975; Mayr, 2009).

Bradfordizing liefert folgende unmittelbare Mehrwerte für den Nutzer:

- eine alternative Sicht auf Suchergebnisse, die nach Kernzeitschriften (Garfield, 1996; im Fall der Monographien auf Kernverlage) umorganisiert sind,
- eine alternative Sicht auf Publikationsquellen innerhalb eines Suchraums, die intuitiv näher am Forschungsprozess liegt als textstatistische Verfahren (z. B. best match) oder traditionelle boolesche Verfahren (exact match),
- eine vermutlich höhere fachliche Relevanz (topicality) der Dokumente nach dem Re-Ranking.

Aufgrund der Robustheit und Allgemeingültigkeit des BLS für die Verteilung von Forschungsliteratur ist davon auszugehen, dass das Bradfordizing gerade in föderierten und interdisziplinären Suchumgebungen mit unterschiedlichen Informationsbeständen praktisch operabel und gewinnbringend ist.

In der Studie (Mayr, 2009) wird Bradfordizing auf typische Information Retrieval Topics (Fragestellungen) und die beiden Dokumenttypen Zeitschriftenartikel und Monographien angewendet. Die Dokumente zu diesen Fragestellungen stammen aus unterschiedlichen Datenbanken und wurden alle intellektuell bzgl. ihrer Relevanz zur Ausgangsfragestellung (Topic) bewertet. Da die Zeitschriftenartikel i. d. R. eine identifizierende Nummer der Zeitschrift (ISSN-Nummer) tragen, kann die anschließende Häufigkeitsanalyse (das eigentliche Bradfordizing) auf Basis der ISSN (International Standard Serial Number) erfolgen. Das gleiche Verfahren kann auf die ISBN-Nummer bei den Monographien angewendet werden. Die ISBN (International Standard Book Number) ist ein Identifier für Monographien und andere selbstständige Veröffentlichungen, in dem der Verlag, der die Publikation herausgibt, kodiert ist. Die Verlagsnummer ist eine ein- oder mehrstellige Ziffer, die eindeutig einem Verlag zugeordnet ist. Dieser Verlags-Code wurde für die Häufigkeitsanalyse (Bradfordizing) verwendet.

Für jedes Topic und jeden Dokumenttyp lassen sich aus den Bewertungsdaten die Anteile der relevanten bzw. nichtrelevanten Dokumente für die jeweilige Bradford-Zone bestimmen. Daraus lassen sich Precision-Werte berechnen. Die Precision misst das Verhältnis von gefundenen relevanten Dokumenten zu einer Suchanfrage an allen gefundenen Dokumenten aus einer Kollektion. Die Precision misst damit die Präzision bzw. Exaktheit eines Retrieval-Ergebnisses. Wenn nur relevante Dokumente aus der Kollektion zu einer Fragestellung gefunden werden, nimmt die Precision den Wert 1 an.

Zusammenfassung

Folgende beide Thesen können als Ergebnisthesen formuliert werden:

1. Die Anwendung des Bradfordizings bzw. das Re-Ranking nach Kernzeitschriften für thematische Dokumentmengen führt zu signifikanten Verbesserungen der Precision zwischen den drei Zonen (Core, Zone 2 und Zone 3). Die Kernzeitschriften (Core) beinhalten signifikant mehr relevante Dokumente als Zeitschriften der Zone 2 oder der Zone 3. Der größte Precision-Gewinn ergibt sich zwischen Core und Zone 3-Zeitschriften.
2. Bradfordizing für thematisch konzentrierte Dokumentmengen lässt sich erfolgreich auf Monographien (Verlag als Selektionskriterium) übertragen. Die Anwendung des Bradfordizings bzw. das Re-Ranking nach Verlagen für Monographien führt zu geringeren Verbesserungen der Precision zwischen den drei Zonen (Core, Zone 2 und Zone 3). Der größte Precision-Gewinn ergibt sich auch hier wie bei den Zeitschriften zwischen Core und Zone 3-Verlagen.

Die Ergebnisse dieser Arbeit gehen in das DFG-Projekt „Retrieval-Mehrwertdienste zur Weiterentwicklung wissenschaftlicher Fachportale wie vascoda und sowiport. Suchexpandierung und Re-Ranking“ (kurz IR-Mehrwertdienste) ein, das im Januar 2009 gestartet wurde (siehe dazu <http://www.gesis.org/index.php?id=2479>).

Literatur

1. Bradford, Samuel C. (1948): Documentation. London: Lockwood. 156 S.
2. Garfield, Eugene (1996): The Significant Scientific Literature Appears In A Small Core Of Journals. In: The Scientist 10, Nr. 17, S. 13. URL: [http://www.garfield.library.upenn.edu/commentaries/tsv10\(17\)p13y090296.html](http://www.garfield.library.upenn.edu/commentaries/tsv10(17)p13y090296.html)
3. Mayr, Philipp (2009): Re-Ranking auf Basis von Bradfordizing für die verteilte Suche in Digitalen Bibliotheken. Philosophische Fakultät I, Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin. 237 S., URL: <http://edoc.hu-berlin.de/dissertationen/mayr-philipp-2009-02-18/PDF/mayr.pdf>
4. Mayr, Philipp; Mutschke, Peter; Petras, Vivien (2008): Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking. In: Library Review 57, Nr. 3, S. 213-224. URL: http://www.ib.hu-berlin.de/~mayr/arbeiten/mayr-et al_LR08.pdf
5. Mutschke, Peter (2004): Autorennetzwerke: Netzwerkanalyse als Mehrwertdienst für Informationssysteme. S. 141-162. In: Bekavac, Bernard; Herget, Josef; Rittberger, Marc (Hrsg.): Information zwischen Kultur und Marktwirtschaft. Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004).
6. Umstätter, Walther (2005): Anmerkungen zu Birger Hjørland und Jeppe Nicolaisen: Bradford's Law of Scattering: Ambiguities in the Concept of "Subject". In: LIBREAS, Nr. 3. URL: http://www.ib.hu-berlin.de/~libreas/libreas_neu/ausgabe3/008ums.htm
7. White, Howard D. (1981): 'Bradfordizing' search output: how it would help online users. In: Online Review 5, Nr. 1, S. 47-54
8. Worthen, D. B. (1975): The application of Bradford's law to monographs. In: Journal of Documentation 31, Nr. 1, S. 19-25