Hans-Christoph Hobohm (Hrsg.)

Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten

Information Science between Virtual Infrastructure and Material Lifeworlds

Unter Mitarbeit von Judith Pfeffing

Proceedings des 13. Internationalen Symposiums
für Informationswissenschaft (ISI 2013)

Potsdam, 19. bis 22. März 2013

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Verwendung von Skalenbewertungen in der Evaluierung von Web-Suchmaschinen

Dirk Lewandowski

Hochschule für Angewandte Wissenschaften Hamburg
Department Information
Finkenau 35, D-22081 Hamburg
dirk.lewandowski@haw-hamburg.de

Zusammenfassung

Ein Nachteil von Retrievaleffektivitätsstudien liegt darin, dass in der Regel nur geringe Mengen von Suchanfragen verwendet werden. Weiterhin werden Relevanzbewertungen meist nur binär angegeben. In der vorliegenden Studie wird anhand der Ergebnisse der Suchmaschinen Google und Bing zu 1.000 Suchanfragen der Unterschied zwischen binären Bewertungen und Skalenbewertungen untersucht. Es zeigt sich, dass sich keine gravierenden Unterschiede in dem Vergleich der Suchmaschinen ergeben. Es wird empfohlen, Skalenbewertungen einzusetzen, da sie die Relevanzbewertungen der Juroren genauer abbilden können.

Abstract

In this study we compare binary relevance assessments with assessments made on a 5-point scale using results from the search engines Bing and Google. We use 1,000 queries and the first 10 results of each query. The results show that there are no significant differences between the two different methods of assessment. Nevertheless, we recommend using 5-point scales as they allow for a more differentiated view of relevance assessment.

In: H.-C. Hobohm (Hrsg.). Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten. Tagungsband des 13. Internationalen Symposiums für Informationswissenschaft (ISI 2013), Potsdam, 19.—22. März 2013. Glückstadt: Verlag Werner Hülsbusch, 339—348.

1 Einleitung

Suchmaschinen sind weiter im Interesse der informationswissenschaftlichen Forschung. Vor allem die Adaption von Tests zur Retrievaleffektivität auf Suchmaschinen wird bearbeitet, und in den vergangenen Jahren wurden bereits zahlreiche Tests durchgeführt. Allerdings lässt sich die Methodik aus Information-Retrieval-Studien nicht eins zu eins auf Suchmaschinen übertragen. Zum Beispiel spielen bei den Web-Suchmaschinen die Präsentation der Ergebnisseiten und die Trefferbeschreibungen eine entscheidende Rolle.

Auf der anderen Seite ist zu fragen, inwieweit Erkenntnisse aus der IR-Evaluierung auch für den Suchmaschinenbereich Gültigkeit beanspruchen können. Während die informationswissenschaftliche Forschung etwa das Thema Skalenbewertungen eingehend behandelt hat (und zumindest für die TREC-Evaluierungen zu dem Schluss gekommen ist, dass binäre Bewertungen das Nutzungsmodell hinreichend abbilden), ist noch unklar, ob binäre Bewertungen auch im Kontext der Websuche ausreichend sind.

Die dieser Studie zugrundeliegende Forschungsfrage lautet daher: „Inwieweit spielt es bei der Bewertung der Relevanz von Suchmaschinen-Treffern eine Rolle, ob die Bewertung binär oder auf einer Skala erfolgt?“

Ziel der Untersuchung war es, Suchmaschinen hinsichtlich ihrer Fähigkeit zu testen, relevante Treffer zu einer Vielzahl von Informationsbedürfnissen zu finden.

2 Bisherige Arbeiten

Die Evaluierung von Information-Retrieval-Systemen ist ein Kernbereich der Informationswissenschaft. Seit Jahrzehnten werden Methoden entwickelt, um die Trefferqualität von IR-Systemen zu bewerten und zu vergleichen. Dabei ist zu unterscheiden zwischen Evaluierungen auf Basis von Testkollektionen und solchen, die bestehende IR-Systeme ohne internen Zugang zu diesen Vergleichen. Der Vorteil ersterer Methode liegt in den besser zu kontrollierenden Testbedingungen, während der große Nachteil darin liegt, dass die Übertragbarkeit der Ergebnisse auf größere bzw. andere Kollektionen strittig ist und vor allem, dass nur Systeme getestet werden können, die tatsächlich

an der jeweiligen Evaluierungsrunde teilnehmen. Dies ist gerade bei Web-Suchmaschinen nicht der Fall.

Dem gegenüber steht allerdings die enorme Bedeutung der Web-Suchmaschinen für die Beschaffung von Informationen und das öffentlich Interesse an Informationen über ihre Qualität und Tauglichkeit. So wurden in den letzten Jahren zahlreiche Studien durchgeführt, deren Ziel es war, die Retrievaleffektivität von Suchmaschinen im Web zu messen (vgl. u.a. Griesbaum 2004; Lewandowski 2008; Tawileh/Griesbaum/Mandl 2010; Véronis 2006).

Zur Unterscheidung zwischen binären Bewertungen und Skalenbewertungen wurden einige Studien durchgeführt (u.a. Greisdorf/Spink 2001), allerdings beziehen sich diese nicht explizit auf die Websuche. Es ist aber davon auszugehen, dass aufgrund der vielen Suchanfragen, die in der Websuche von den Suchenden selbst als nicht allzu bedeutend eingeschätzt werden, die Schwelle zwischen binären Bewertungen und Skalenbewertungen recht niedrig liegt, d.h. Dokumente schnell als relevant angesehen werden.[1]

3 Testaufbau

Die für den Test verwendeten Suchanfragen entstammen den Logfiles der Suchmaschine auf dem T-Online-Portal. Im Rahmen des Portals wird eine Websuche angeboten, die Ergebnisse von Google anzeigt. T-Online wird stark genutzt und bedient in Deutschland ca. vier Prozent der an allgemeine Suchmaschinen gestellten Suchanfragen (Webhits 2012). Während die Informationsbedürfnisse der Nutzer unterschiedlicher Suchmaschinen natürlich weit auseinander liegen können, ist gerade bei Suchsystemen, die sich an die Allgemeinheit wenden und eine sehr große Anzahl von Suchanfragen bedienen, davon auszugehen, dass sich die dort eingegebenen Suchanfragen nicht grundlegend unterscheiden, auch wenn aufgrund der Nicht-Verfügbarkeit entsprechender Daten von anderen Suchmaschinen ein empirischer Vergleich nicht möglich ist.

1 Diese Hypothese wurde aus einer Analyse der Skalenbewertungen aus einer Vorgängerstudie (Lewandowski 2008) abgeleitet.

Aus einem Set von vielen Mio. Suchanfragen[2] wurde ein Zufallsstichprobe von 1.000 Suchanfragen gezogen, zu denen anschließend jeweils die ersten zehn Treffer der Suchmaschinen Google und Bing erfasst und hinsichtlich ihrer Relevanz von Juroren bewertet wurden.

Das Sample wurde gebildet, indem zuerst aus den Suchanfragen zehn Gruppen gebildet wurden, die jeweils zehn Prozent des Suchvolumens im untersuchten Zeitraum auf sich vereinen. Da die Verteilung der Suchanfragen stark linksschief ist, hätte bei einer direkten Zufallsauswahl die Gefahr bestanden, dass wenig populäre Suchanfragen bevorzugt berücksichtigt worden wären, während populäre Suchanfragen, deren Beantwortung für die Beurteilung einer Suchmaschine eine wichtige Rolle spielen, ggf. unter den Tisch gefallen wären.

Aus jeder Gruppe wurden 360 Suchanfragen zufällig ausgewählt und einer Jurorin zur Klassifizierung hinsichtlich des Anfragetyps (informationsorientiert, navigationsorientiert, transaktionsorientiert) vorgelegt. Nur informationsorientierte Suchanfragen gingen in die weitere Auswertung ein. Aus den ersten Gruppen (mit den sehr populären Suchanfragen) wurden alle informationsorientierten Suchanfragen verwendet (bei den populären Suchanfragen ist die Gesamtgröße der Gruppe recht gering und der Anteil der navigationsorientierten sehr hoch, weshalb das Ziel von 100 informationsorientierten Suchanfragen je Gruppe hier nicht erreicht werden konnte). Aus den anderen Gruppen wurden jeweils gleich viele Suchanfragen ausgewählt, so dass sich insgesamt eine Menge von 1.000 Suchanfragen ergab.

Die Erfassung der Suchergebnisse erfolgte mit dem Relevance Assessment Tool (RAT) (vgl. Lewandowski/Sünkler 2012). RAT ermöglicht eine automatische Erfassung von Suchmaschinen-Ergebnisseiten und den Ergebnisdokumenten selbst, welche in einer Datenbank gespeichert werden, so dass die Bewertung nicht von der unveränderten Verfügbarkeit der URLs zum Zeitpunkt der Bewertung abhängt. Die Suchanfragen wurden in RAT hochgeladen und die Ergebnisse der beiden Suchmaschinen automatisch in die Datenbank geschrieben.

Auch die Bewertung durch die Juroren erfolgte mittels RAT. Mit der Software lassen sich beliebige Fragen zu den Ergebnisdokumenten stellen; die Bewertung lässt sich mittels Crowdsouring beliebig verteilen. Im aktuellen Fall wurde der Zugangscode zur Studie über den Studentenverteiler des

2 Die genaue Zahl kann hier aus Wettbewerbsgründen nicht genannt werden.

Departments Information der HAW Hamburg und über Facebook verteilt. Bei erfolgreicher Bewertung aller Dokumente zu einer Aufgabe gab das System automatisch einen Amazon-Gutschein über fünf Euro aus. Diese kleine Belohnung motivierte die Juroren ungemein, so dass die Bewertung aller Dokumenten innerhalb einiger Stunden bewerkstelligt werden konnte.

Die Juroren bekamen die Dokumente zu einer Suchanfrage in zufällig gemischter Reihenfolge und ohne Hinweise auf die Suchmaschinen ausgegeben. Sie wurden um zwei Bewertungen gebeten: einerseits um eine binäre Relevanzbewertung (ist relevant / ist nicht relevant), andererseits um eine differenzierte Bewertung auf einer Skala („Wie relevant ist dieses Dokument?“, von 0 bis 4). Dokumente, die von beiden Suchmaschinen ausgegeben wurden, wurden den Juroren nur einmal zur Bewertung vorgelegt und vom System automatisch wieder den entsprechenden Suchmaschinen zugeordnet.

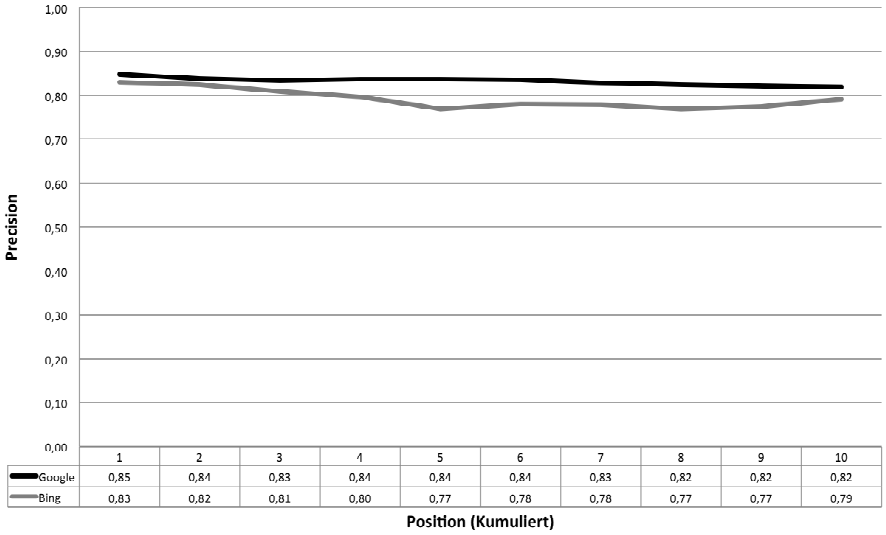
Den Juroren wurden neben der Suchanfrage keine weiteren Informationen zum Informationsbedürfnis vorgelegt. Dies geschah einerseits aufgrund des Aufwands, den es erfordert hätte, um zu jeder Suchanfragen ein nicht beliebiges Informationsbedürfnis zu konzentrieren (etwa durch ein Verfahren, wie es Huffman/Hochster (2007) angewendet haben), andererseits aufgrund der Schwierigkeit, aus ‚nackten‘ Suchanfragen überhaupt auf die Informationsbedürfnisse schließen zu können (vgl. Lewandowski/Drechsler/Mach 2012). So wurde diese Aufgabe den Juroren überlassen, die die vorgegebene Suchanfrage beliebig in ihrem Sinne interpretieren konnten. Da die Dokumente beider Suchmaschinen zu einer Suchanfrage von einem Juror bewertet wurden, ist selbst bei einer eher abseitigen Interpretation keine Verzerrung im Vergleich der Suchmaschinen zu konstatieren, sondern allenfalls in der Gesamtbewertung der Suchmaschinen. Dabei ist davon auszugehen, dass beide Suchmaschinen in diesem Test eher besser abgeschnitten haben als dies der Fall gewesen wäre, wenn vorgegebene Informationsbedürfnisse verwendet worden wären.

4 Ergebnisse

In diesem Abschnitt werden die Ergebnisse, getrennt nach binären und Skalenbewertungen, dargestellt.

4.1 Binäre Bewertungen

In die Untersuchung gingen insgesamt 9.265 (Google) bzw. 9.790 (Bing) binäre Bewertungen ein. Der Anteil der als relevant bewerteten Treffer betrug für Google 82,0 Prozent, bei Bing 79,3 Prozent. Betrachtet man den Precision-Graphen in Abbildung 1, so zeigt sich, dass die Suchmaschinen auch bei der Berücksichtigung der Trefferposition nahe beieinander liegen.



	1	2	3	4	5	6	7	8	9	10
Google	0,85	0,84	0,83	0,84	0,84	0,84	0,83	0,82	0,82	0,82
Bing	0,83	0,82	0,81	0,80	0,77	0,78	0,78	0,77	0,77	0,79

Abb. 1 Precision-Graph (binäre Bewertungen)

4.2 Skalenbewertungen

Wie in Abbildung 2 zu sehen ist, verteilen sich die Skalenbewertungen bei den beiden Suchmaschinen recht ähnlich, wobei deutlich wird, dass der Anteil der als völlig irrelevant (auf der Skala mit 0 bewertete Treffer) bei Bing deutlich höher ist.

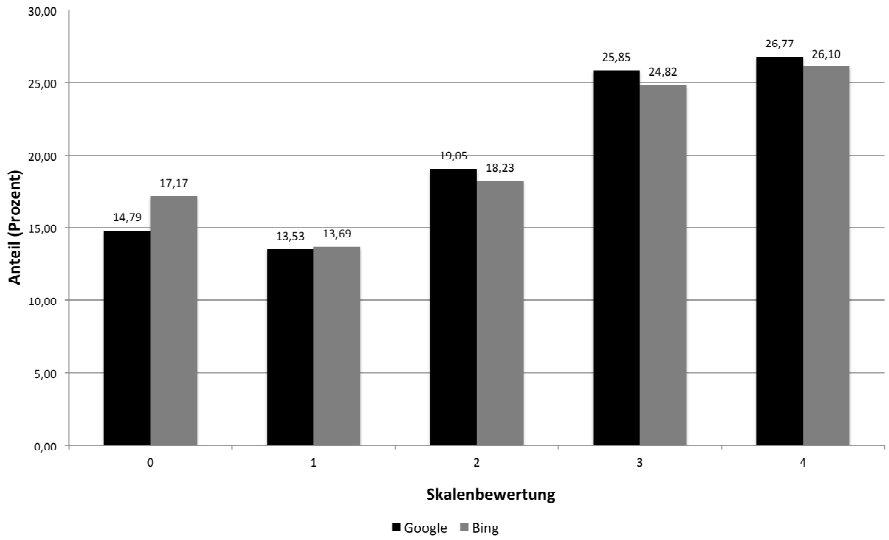


Abb. 2 Verteilung der Skalenbewertungen

In Abbildung 3 ist der Precision-Graph nur für die top bewerteten Treffer dargestellt; es werden also nur die mit 4 auf der Skala bewerteten Treffer dargestellt. Auch hier zeigen sich keine größeren Unterschiede zwischen den Suchmaschinen.

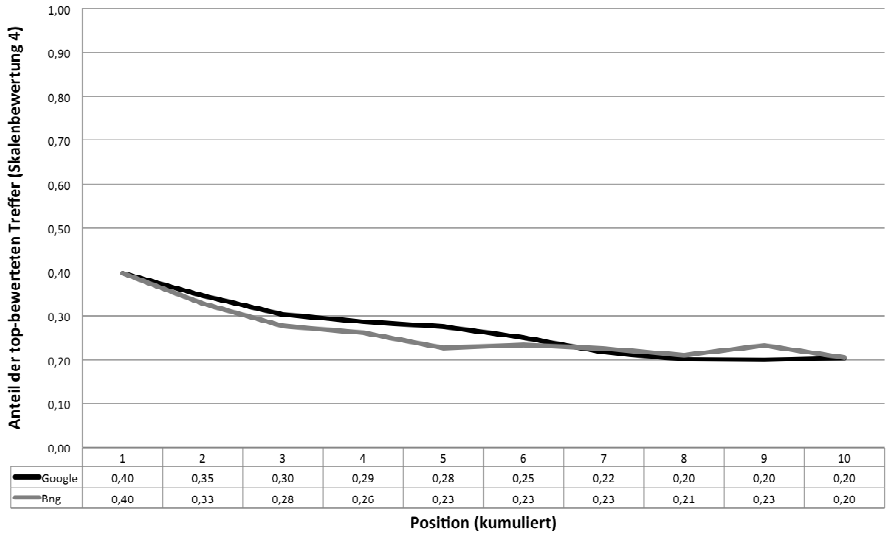


Abb. 3 Anteil der Treffer mit Skalenbewertung 4

5 Diskussion

Aus dem Vergleich zwischen binärer Bewertung und Skalenbewertung ergibt sich wider Erwarten kein gravierender Unterschied beim Vergleich der beiden Suchmaschinen. Es zeigt sich keine klare Überlegenheit der Skalenbewertungen gegenüber den binären Bewertungen, jedoch empfehlen wir trotzdem den Einsatz von Skalenbewertungen, da sie eine differenziertere Bewertung ermöglichen. Zu empfehlen ist auch die parallele Abfrage beider Bewertungen (wie in der vorliegenden Untersuchung).

Allerdings unterliegen die Ergebnisse dieser Studie einigen Einschränkungen:

- Den Juroren wurden in dieser Untersuchung nur die ‚nackten' Suchanfragen vorgelegt, aus denen sich in vielen Fällen nicht das exakte Informationsbedürfnis der Nutzer, die die Suchanfrage ursprünglich gestellt haben, ableiten lässt. Allerdings ist gerade die Definition von Informationsbedürfnissen aus Suchanfragen problematisch, da unter Umständen falsch interpretiert wird und die Untersuchung damit eine Verzerrung hin zu von den Testleitern gemachten Interpretationen erhält. Auf der anderen Seite sind beide im Test behandelten Suchmaschinen von den Interpretationen durch die Juroren gleichermaßen betroffen, so dass sich eventuelle Verzerrungen ausgleichen. Allerdings ist zu bedenken, dass die Bewertungen bei nicht genau definierten Informationsbedürfnissen in der Regel positiver ausfallen, weshalb die in dieser Studie gemachten Relevanzbewertungen positiver sein dürften als in einer natürlichen Recherchesituation.
- Die Auswahl der Juroren wurde in dieser Untersuchung nicht kontrolliert, sondern es wurde ein Zugangscode an alle Juroren ausgegeben mit der Möglichkeit, diesen weiter zu verbreiten. Da noch immer ungeklärt ist, inwieweit sich die Urteile von Juroren unterschiedlicher Gruppen im Kontext der Web-Suche voneinander unterscheiden, ist es durchaus möglich, dass, falls andere Juroren rekrutiert worden wären, auch andere Bewertungen herausgekommen wären. Allerdings ist auch hier anzumerken, dass sich vielleicht das grundsätzliche Bewertungsniveau verändert hätte, die Auswahl allerdings keine Auswirkungen auf den Vergleich zwischen den Suchmaschinen gehabt hätte.
- Zuletzt sei anzumerken, dass in dieser Untersuchung die Ergebnisse der Universal Search nicht mit berücksichtigt wurden. Da Suchmaschinen

diese Ergebnisse auf unterschiedliche Weise und in unterschiedlichem Umfang einbinden, können sich Verzerrungen in der Untersuchung ergeben. Ebenso ist zu bedenken, dass, wenn eine URL bereits als Universal-Search-Ergebnis angezeigt wird, diese aus den regulären Treffern herausgenommen wird. Dadurch ist es möglich, dass in dieser Untersuchung wichtige Ergebnisse fehlen. Hier lässt sich aufgrund der Forschungslage nicht entscheiden, ob eine Suchmaschine durch diesen Umstand in der Untersuchung benachteiligt wurde.

6 Fazit

Auf der Basis einer umfassenden Studie mit 1.000 Suchanfragen, die zu insgesamt mehr als 19.000 Relevanzbewertungen führten, wurden die Unterschiede von binären Bewertungen und Skalenbewertungen bei der Beurteilung von Suchmaschinentreffern untersucht. Es zeigte sich, dass die Skalenbewertungen beim vorliegenden Testdesign – die Informationsbedürfnisse waren von den Juroren frei aus den vorliegenden Suchanfragen zu konstruieren – keine gravierenden Unterschiede zwischen den beiden Bewertungsarten festgestellt werden konnten. Dennoch wird die Erhebung von Relevanzbeurteilungen auf einer Skala aus theoretischen Erwägungen empfohlen.

Neben den Ergebnissen zu den Skalenbewertungen fielen in dieser Studie auch Daten an, die einen Vergleich der Trefferqualität der beiden bekannten Suchmaschinen Google und Bing anhand eines repräsentativen Samples von Suchanfragen möglich machen. Die beiden im Test behandelten Suchmaschinen haben vergleichbar abgeschnitten. Dies deutet wieder einmal darauf hin, dass die starke Präferenz der Nutzer für Google sich nicht aus der überlegenen Qualität der Treffer dieser Suchmaschine ableiten lässt. Auf der anderen Seite konnte gezeigt werden, dass es mit Bing durchaus eine ernstzunehmende Alternative zu Google gibt, auch wenn die Marktanteile der beiden Suchmaschinen anderes suggerieren.

Literaturverzeichnis

Greisdorf, H.; Spink, A. (2001). Median measure: an approach to IR systems evaluation. In: Information Processing & Management 37 (6), 843–857.

Griesbaum, J. (2004). Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. In: Information Research 9 (4), 1–35. http://informationr.net/ir/9-4/paper189.html <21.1.2013>.

Huffman, S. B.; Hochster, M. (2007). How well does result relevance predict session satisfaction? In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 567–574.

Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. In: Journal of Documentation 64 (6), 915–937.

Lewandowski, D.; Drechsler, J.; Mach, S. von (2012). Deriving Query Intents From Web Search Engine Queries. In: Journal of the American Society for Information Science and Technology 63 (9), 1773–1788.

Lewandowski, D.; Sünkler, S. (2012). Relevance Assessment Tool. Ein Werkzeug zum Design von Retrievaltests sowie zur weitgehend automatisierten Erfassung, Aufbereitung und Auswertung von Daten. In: Ockenfeld, M.; Weller, K.; Peters, I. (Hrsg.). Social Media und Web Science: Das Web als Lebensraum. Proceedings der 2. DGI-Konferenz, 237–249.

Tawileh, W.; Griesbaum, J.; Mandl, T. (2010). Evaluation of five web search engines in Arabic language. In: Atzmüller, M.; Benz, D.; Hotho, A.; Stumme, G. (Hrsg.). Proceedings of LWA2010. Kassel, Germany. http://www.kde.cs.uni-kassel.de/conf/lwa10/papers/ir1.pdf <21.01.2013>.

Véronis, J. (2006). A comparative study of six search engines. Université de Provence. http://sites.univ-provence.fr/veronis/pdf/2006-comparative-study.pdf <21.01.2013>.

Webhits (2012). Webhits Web-Barometer. http://www.webhits.de/deutsch/index.shtml?webstats.html <21.01.2013>.