Investigating the Effects of Popularity Data on Predictive Relevance Judgments in Academic Search Systems

Christiane Behnert
Department of Information
Hamburg University of Applied Sciences
Hamburg, Germany
christiane.behnert@haw-hamburg.de

ABSTRACT

The elements of a surrogate serve as clues to relevance. They may be seen as operationalized relevance criteria by which users judge the relevance of a search result according to their information need. In addition to short textual summaries, today's academic search systems integrate additional data into their search results presentation, for example, the number of citations or the number of downloads. This kind of data can be described as popularity data, serving as factors also incorporated in search engines' ranking algorithms. Past research shows that there are diverse criteria and factors involved in relevance judgements from the user perspective. However, previous empirical studies on relevance criteria and clues examined surrogates that did not include popularity data. The goal of my doctoral research is to gain significant knowledge on the criteria by which users in an academic search situation make relevance judgements based on surrogates that include popularity data. This paper describes the current state of the experimental research design and method of data collection.

CCS CONCEPTS

• Human-centered computing → User models; User studies; Empirical studies in HCI

KEYWORDS

Relevance behavior; predictive judgments; experimental design

ACM Reference format:

Christiane Behnert. 2019. Investigating the effects of popularity data on predictive relevance judgments in academic search systems. In *Proceedings of ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR'19) March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 4 pages.
DOI: https://doi.org/10.1145/3295750.3298978

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03.
https://doi.org/10.1145/3295750.3298978

1 Research problem

The standard elements of a surrogate provide general information about a document retrieved by an Information Retrieval (IR) system. Today's academic search systems integrate additional data into their search results presentation. For example, Google Scholar's results include the number of citations or other versions of the particular work, and the ACM Digital Library provides the number of downloads for search results. This kind of data can be described as popularity data. Popularity data are factors for popularity based on the wisdom of crowds principle [24]. For instance, the more users judge a search result as relevant, the more popular the search result would be. Such popularity factors are also integrated into relevance ranking algorithms [17]. In the library context, popularity data would be, for example, the number of circulations, or the number of copies held by the library [5].

The elements of a surrogate (i.e., metadata and additional information such as popularity data) serve as clues to relevance since they provide a significant proportion of the information by which users judge the relevance of the result according to their information need. These relevance clues build the foundation for the criteria that users employ in the process of relevance judgment, that is, clues may be seen as operationalized relevance criteria. Past research shows that there are diverse criteria and factors involved in relevance judgments from the user perspective [e.g., 3, 13, 20, 21]. Empirical studies on relevance criteria and clues are based on search results that, at that time, usually did not include additional data in general or popularity data in particular. The goal of my doctoral research is to gain significant knowledge about the criteria by which users make relevance judgments of surrogates when the surrogates include popularity data. The focus is on academic search systems and user behavior in an academic search situation.

2 Related research

The related work is synthesized according to three topics: (a) relevance criteria and clues to relevance, (b) document representations as objects of investigations in relevance criteria studies, and (c) user models of relevance criteria.

2.1 Relevance criteria

In the 1990's, a shift from a system-oriented view towards a user-oriented view on relevance in IR occurred, which is mirrored by studies on relevance criteria that were undertaken at this time (see [14]). Saracevic [22] synthesizes what we have learned from decades of relevance research. He provides an overview of twenty-one "observational, empirical, or experimental" studies on relevance clues and criteria, including the influential work by Barry and Schamber [3]. They had analyzed criteria according to which users judge the relevance of a document. Among other studies [4, 8, 10, 11, 16, 25], their results show that several criteria beyond topical relevance are involved in relevance judgments, for example, validity, recency, availability, and credibility of the information source.

With respect to the dynamic nature of the Web and its exponential growth, credibility and quality are both significant factors to filter and judge information retrieved by Web search engines [20]. Credibility, in particular, can be considered in terms of cognitive authority [31], which is highly subjective: A person is not only an expert but a cognitive authority as well, when his or her statements of knowledge are accepted by others as truth – the information is trustworthy – while he or she would also influence other people's thoughts. Assessments of an author's cognitive authority are based on his or her present reputation and accomplishments [18]. Thus, in an academic context, information about an author's impact is helpful for users making relevance judgments. A similar finding was reported by Wang and Soergel [30], as their results suggest a need for providing information about the status or impact of authors and publishers as part of the search results' presentations.

2.2 Document representations as objects of investigation in studies on relevance criteria

Rieh [19] distinguishes between predictive and evaluative relevance judgments within the information search process, i.e., judgments based on the search result and judgments made based on the full-text document. Past studies on relevance criteria involving predictive relevance judgments are, for example, the study on document representations and clues by Barry [2], the study on document information elements by Wang [29] and Wang and Soergel [30], and the study on individual criteria in the process of judging abstracts by Tang & Solomon [26]. Overall findings suggest that titles and abstracts provide most of the relevance clues [22], which confirms the great importance of topicality and content-related criteria.

In Web search, snippets and hyperlinks may be seen as website surrogates. For example, Savolainen and Kari [23] conducted a study on selection behavior based on hyperlinks. Rieh [19] investigated how users decide on what website to look. Her study focused on information quality and cognitive authority as factors influencing web search behavior. In another study by Papaeconomou, Zijlema and Ingwersen [15], participants' eye movements were recorded to explore the "relevance hot spots" of Web pages applied during the judgment process, while Balatsoukas and Ruthven [1] investigate relevance criteria in judgment process on Google also applying an eye-tracking approach. However, those studies were not set in an academic search context, i.e., did not investigate predictive judgments in academic search systems.

As mentioned above, today's academic web search engines and other academic information systems integrate popularity data into their search results representations. At the time of the previous research reported above, such data were not parts of the search results, or did not focus on an academic context. However, more recently published research on relevance criteria that involve surrogates in academic search systems as objects of investigation [6, 27, 28] lack enrichment with additional data such as popularity data.

To the best of my knowledge, no studies on the relevance of surrogates containing this type of data in an academic context, while at the same time taking into account the user perspective, have been published to date.

2.3 User models on relevance criteria

Saracevic [22] states that "All IR and human information behavior models have relevance at their base either explicitly or as an invisible hand – in effect they are relevance models" (p. 40). Although these models may incorporate relevance and relevance judgments, they do not demonstrate the criteria and factors involved in the judgment process. Models that intend to do so are (a) the Document selection model for academic users in agricultural economics [30], initially developed by Wang [29], and (b) the Model of judgment of information quality and cognitive authority by Rieh [19]. While Wang studied the cognitive processes and document information elements that affect users in their decision making, Rieh's model focuses on two criteria particularly salient in Web search, namely, information quality, and cognitive authority.

Again, additional or popularity data as described above were not elements of the search results that the participants in their studies had judged.

To sum up, the results of the literature review revealed a gap in research on relevance criteria employed by users of academic search systems with regard to popularity data.

3 Research questions

The overall purpose of this research is to gain significant knowledge on the criteria by which users of academic search systems judge the relevance of a search result. The goal is to develop a user model on predictive relevance judgments in academic information systems. An initial version of this user model has been developed based on the results of the literature review, which helped me to narrow my research focus onto relevance criteria based on popularity data and, thus, to concretize my research questions. The model explicitly distinguishes between relevance clues, relevance criteria and relevance factors and, thus, offers a systematic representation of the influencing variables on the predictive relevance judgment process in academic search systems. In this model, popularity

data are presented as dynamic, user-independent elements of a surrogate because they bear information that may change over time and are independent of the user in his or her particular information search situation. Following this definition, the research questions were formulated as follows:

RQ I: Which dynamic, user-independent elements of the surrogate influence how people searching for information judge the relevance of a search result?

RQ II: Which of these elements influence the relevance decision to what extent?

RQ III: How are the criteria derived from these elements weighted against each other?

4 Methodology

The research goal is pursued in five steps: (1) an extensive literature review of related research, (2) the development of an initial user model of relevance criteria which allows to formulate hypotheses, (3) an empirical examination through an online experiment with human test persons, (4) the statistical data analysis, (5) the alteration of the initial user model after reflecting the results of the online experiment.

The research questions will be addressed by conducting a user study. The study applies an experimental design, where participants are asked to perform predictive relevance judgments of search results to given search tasks.

At the time of writing this proposal, I have designed the experiment with a first version of the data collection instrument, as described in the following sections.

4.1 Experimental research design

Much research on the concept of relevance, as well as the factors and criteria that may influence users' relevance judgments has been undertaken [e.g., 13, 21]. However, Buckland [9], claims relevance to be immeasurable "in the normative sense of formal and physical sciences such as mathematics and physics, based on formal conjecture and refutation" (p. 163), because of its subjective and dynamic nature. He argues that a relevant document "must be useful to an actual human being's mental activity" (p. 161).

One approach to studying relevance with regard to human behavior is to develop an experimental research design, as common in social science disciplines, e.g., psychology and education, where experiments are frequently conducted to study human behavior. Information Behavior and Interactive Information Retrieval research may strongly benefit from experiments to gain significant insights into search behavior and interaction between users and IR systems and to get more valid and reliable results [13].

In an experiment, a researcher aims to establish causality between variables (stimuli and effects). To test whether a causal conclusion can be derived, some manipulation of variables (stimuli) and control of conditions (e.g., randomization, counter-balancing) are required. Through this, variables can be isolated and studied systematically [12]. Thus, the factors or criteria involved in relevance judgments, i.e., variables, can be isolated and modelled, which in real-life situations would not be feasible.

Table 1: Levels of independent variables (IVs)

IV	Level A	Level B	Level C
No. of citations (work)	Low	High	Not provided
No. of citations (author)	Low	High	Not provided
Author impact	Low	High	Not provided
No. of downloads	Low	High	Not provided

To investigate the effects of popularity data (independent variables) on academic users' predictive relevance judgments (dependent variable), I developed a multifactorial within-subjects design. Since we assume that a variety of clues affects relevance judgments, a multifactorial design is needed to test a combination of factors. Table 1 provides an overview of the independent variables (IV). Each of the IV will be manipulated on three levels: (a) a low number, (b) a high number, (c) no information will be provided, leading to 81 experimental conditions (3x3x3x3=81).

4.2 Data collection

The online experiment will be conducted using the online survey tool EFS Survey (by Questback). Aside from a questionnaire to collect demographic data and information on search self-efficacy as well as experiences with academic search systems, study participants will judge nine search results to a total of nine search tasks (81 surrogates to test all experimental conditions), on a visual analogue scale with a slider from zero to 100. Within the task description, the concept of relevance is operationalized as usefulness, stating: "Please use the slider to judge how useful you think each search result is for answering the search task!"

Both surrogates and search tasks will be presented in a randomized order. The search tasks will be developed based on the concept of simulated work task situations [7]. Topics for search tasks will be information science related topics that participants from different academic backgrounds are assumed to be familiar with in order to understand the given task. Topics include, for example, the peer review process, alternative metrics with regard to social media, and Wikipedia.

So far, I selected search results from topically relevant information science sources for two tasks, i.e., review articles in ARIST and JASIS&T's Advances in Information Science. All search results are manually produced images inspired by the Google Scholar search results presentation design. Participants will not be able to click on links or even view the full text to avoid comparability issues and reducing the internal validity of the experiment.

Participants will be graduate students (Master or PhD level) or academic staff from different academic disciplines. As an incentive and as a thank you for participation, an Amazon gift

voucher à 10 EUR will be sent to the participant after completing the questionnaire via e-mail.

5 Future plans

At the beginning of 2019, I will conduct a pilot study to confirm the validity of the experimental research design. Further, the preliminary results will verify preceding choices of type and number of independent variables, characteristics of the sample, presentation of surrogates, and measures for statistical analysis. If need be, the research design will be revised.

The size of the sample for the study itself is planned to be at least n=400 to get statistically significant and sound results. To establish the exact sample size needed, the software tool G*Power for statistical analysis will be used.

ACKNOWLEDGMENTS

This research is funded by a three-year Ph.D. scholarship from the Hamburg University of Applied Sciences, Germany. I thank my supervisors Prof. Dr. Ulrike Spree, Hamburg University of Applied Sciences, and Prof. Dr. Joachim Griesbaum, University of Hildesheim, Germany, for their ongoing support.

REFERENCES

[1] Balatsoukas, P. and Ruthven, I. 2012. An eye-tracking approach to the analysis of relevance judgments on the Web: The case of Google search engine. *Journal of the American Society for Information Science and Technology*. 63, 9 (Sep. 2012), 1728–1746. DOI:https://doi.org/10.1002/asi.22707.

[2] Barry, C.L. 1998. Document representations and clues to document relevance. *Journal of the American Society for Information Science*. 49, 14 (Jan. 1998), 1293–1303. DOI:https://doi.org/10.1002/(SICI)1097-4571(1998)49:14<1293::AID-ASI7>3.0.CO;2-E.

[3] Barry, C.L. and Schamber, L. 1998. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*. 34, 2–3 (März 1998), 219–236. DOI:https://doi.org/10.1016/S0306-4573(97)00078-2.

[4] Bateman, J. 1998. Changes in Relevance Criteria: A Longitudinal Study. *Proceedings of the 61st ASIS Annual Meeting* (1998), 23–32.

[5] Behnert, C. and Lewandowski, D. 2015. Ranking search results in library information systems — Considering ranking approaches adapted from web search engines. *The Journal of Academic Librarianship*. 41, 6 (Nov. 2015), 725–735. DOI:https://doi.org/10.1016/j.acalib.2015.07.010.

[6] Beresi, U.C. et. al. 2010. Why did you pick that? Visualising relevance criteria in exploratory search. *International Journal on Digital Libraries*. 11, 2 (June 2010), 59–74. DOI:https://doi.org/10.1007/s00799-011-0067-7.

[7] Borlund, P. und Ingwersen, P. 1997. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*. 53, 3 (Aug. 1997), 225–250. DOI:https://doi.org/10.1108/EUM0000000007198.

[8] Bruce, H.W. 1994. A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science*. 45, 3 (Apr. 1994), 142–148. DOI:https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<142::AID-ASI4>3.0.CO;2-6.

[9] Buckland, M.K. 2017. *Information and society*. MIT Press.

[10] Cool, C. u. a. 1993. Characteristics of texts affecting relevance judgments. *Proceedings of the 14th Annual National Online Meeting, New York, May 4 - 6, 1993* (1993), 77–84.

[11] Howard, D.L. 1994. Pertinence as reflected in personal constructs. *Journal of the American Society for Information Science*. 45, 3 (1994), 172–185. DOI:https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<172::AID-ASI7>3.0.CO;2-V.

[12] Kelly, D. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*. 3, 1—2 (2009). DOI:https://doi.org/10.1561/1500000012.

[13] Kelly, D. and Cresenzi, A. 2016. From design to analysis: Conducting controlled laboratory experiments with users. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '16* (New York, New York, USA, 2016), 1207–1210.

[14] Mizzaro, S. 1997. Relevance: The whole history. *Journal of the American Society for Information Science*. 48, 9 (Sep. 1997), 810–832. DOI:https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U.

[15] Papaeconomou, C. et. al. 2008. Searchers' relevance judgments and criteria in evaluating web pages in a learning style perspective. *Proceedings of the second international symposium on Information interaction in context - IIiX '08* (New York, New York, USA, 2008), 123–132.

[16] Park, T.K. 1993. The nature of relevance in Information Retrieval: An empirical study. *The Library Quarterly*. 63, 3 (July 1993), 318–351.

[17] Plassmeier, K. et. al. 2015. Evaluating popularity data for relevance ranking in library information systems. *Proceedings of the 78th ASIS&T Annual Meeting* (2015).

[18] Rieh, S.Y. 2009. Credibility and cognitive authority of information. *Encyclopedia of Library and Information Sciences*. CRC Press. 1337–1344.

[19] Rieh, S.Y. 2002. Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*. 53, 2 (Jan. 2002), 145–161. DOI:https://doi.org/10.1002/asi.10017.

[20] Rieh, S.Y. and Belkin, N.J. 1998. Understanding judgment of information quality and cognitive authority in the WWW. *Proceedings of the 61st ASIS Annual Meeting* (1998), 279–289.

[21] Saracevic, T. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*. 58, 13 (Nov. 2007), 2126–2144. DOI:https://doi.org/10.1002/asi.20681.

[22] Saracevic, T. 2016. The Notion of relevance in information science: Everybody knows what relevance is. But, what is it really?. Morgan & Claypool.

[23] Savolainen, R. and Kari, J. 2006. User-defined relevance criteria in web searching. *Journal of Documentation*. 62, 6 (Nov. 2006), 685–707. DOI:https://doi.org/10.1108/00220410610714921.

[24] Surowiecki, J. 2005. *The wisdom of crowds*. Anchor Books.

[25] Tang, R. and Solomon, P. 1998. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Processing & Management*. 34, 2–3 (March 1998), 237–256. DOI:https://doi.org/10.1016/S0306-4573(97)00081-2.

[26] Tang, R. and Solomon, P. 2001. Use of relevance criteria across stages of document evaluation: On the complementarity of experimental and naturalistic studies. *Journal of the American Society for Information Science and Technology*. 52, 8 (Jan. 2001), 676–685. DOI:https://doi.org/10.1002/asi.1116.

[27] Taylor, A. 2013. Examination of work task and criteria choices for the relevance judgment process. *Journal of Documentation*. 69, 4 (July 2013), 523–544. DOI:https://doi.org/10.1108/JD-12-2011-0054.

[28] Taylor, A. 2012. User relevance criteria choices and the information search process. *Information Processing & Management*. 48, 1 (Jan. 2012), 136–153. DOI:https://doi.org/10.1016/j.ipm.2011.04.005.

[29] Wang, P. 1994. *A cognitive model of document selection of real users of information retrieval systems*. University of Maryland; College of Library and Information Science.

[30] Wang, P. and Soergel, D. 1998. A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science*. 49, 2 (1998), 115–133. DOI:https://doi.org/10.1002/(SICI)1097-4571(1998)49:2<115::AID-ASI3>3.0.CO;2-1.

[31] Wilson, P. 1983. Second-hand knowledge: An inquiry into cognitive authority. Greenwood Press.