Information Services & Use 33 (2013) 53–59
DOI 10.3233/ISU-130691
IOS Press

Designing search engine retrieval effectiveness tests with RAT

Dirk Lewandowski * and Sebastian Sünkler
Department of Information, Hamburg University of Applied Sciences, Hamburg, Germany
E-mails: {dirk.lewandowski, sebastian.suenkler}@haw-hamburg.de

Abstract. We introduce the Relevance Assessment Tool (RAT), which allows researchers to design complex search engine retrieval effectiveness studies. This Web-based tool consists of different modules that researchers can use to design tests, to collect results from different search engines using a screen-scraping approach, and to collect judgments from a multitude of jurors using a crowdsourcing approach. We designed the software as a Web-based application, which allows for a distributed collection of relevance judgments, and the like. Using the Relevance Assessment Tool allows for much larger search engine studies than previously conducted. To our knowledge, there is no comparable tool that was developed in the academic context. Therefore, no information regarding the design of such software was available to-date.

Keywords: Worldwide Web, search engines, evaluation, tool support, screen-scraping, search engine results pages, crowdsourcing

1. Introduction

Search engine research has become an important area within information science [1,11] and in other disciplines, such as communications [18], media studies [13], and the social sciences in general [3]. Throughout its long tradition of evaluating information systems, one of information science's major concerns is the quality of search engines [12]. While quality is a multifaceted concept, the retrieval effectiveness of search engines is still the core component of all extensive quality evaluations.

Numerous search engine retrieval effectiveness studies have already been conducted [6,8,17]; for an overview see Lewandowski [7]. However, these studies are hardly comparable, as they use different methods. Moreover, for time and effort reasons, they use varying numbers of queries, jurors, etc. To reduce effort, software support for conducting retrieval effectiveness tests is needed. Therefore, we investigated whether those researchers used software tools to support their tests. We found that while some researchers developed rudimentary tools only for the test at hand, other studies used a simple paper-based design. In the latter case, jurors were given printouts of documents and were asked to mark these printouts as either relevant or not.

A problem that all persons conducting such studies face is the vast amount of time and effort required to design the study, finding participants willing to judge the documents, preparing the documents before jurors see them (e.g., removing all information that could identify the individual search engine from which a result originates and mixing the results so that the result order does not have an effect on judgments), allocating the documents to the search engines for data analysis, and finally analyzing the

[*] Corresponding author. E-mail: dirk.lewandowski@haw-hamburg.de.

data. The goal of the *Relevance Assessment Tool* (*RAT*) is to significantly reduce the effort needed in all aforementioned tasks, with the exception of finding the jurors.

We designed the tool to allow a flexible test design and to allow different project administrators to conduct multiple tests at the same time. Also, we tried to make the tool easy to use and robust, because we also intended to allow students to experiment with the tool and design their own studies.

2. Components of rat

In this section, we describe the modules of RAT, following the process diagram depicted in Fig. 1. RAT can be used to design a test, to automatically fetch search engine results through screen scraping, and to collect relevance judgments.

2.1. Test design and project administration

In this section, we describe the many possibilities that RAT offers for designing individual retrieval effectiveness studies. The tool's design choices are based on the authors' experiences with studies conducted in the past [7–9], on recommendations from other studies, and on guidelines for conducting retrieval effectiveness tests [2,4,15].

2.1.1. Search tasks

All studies are based on a collection of search tasks for which results from the search engines should be collected. A search task in RAT can consist of a short description of the task, the search query, and a narrative describing what kinds of documents are relevant.

Project administrators can choose to enter each search task into the system individually or to bulk-upload multiple search tasks into RAT using an Excel spreadsheet. Once they are in the system, search tasks can be reused in further studies.

2.1.2. Scales

RAT allows a project administrator to design individual scales or to use scales already applied in a previous study. Also, the administrator can design individual questions. For example, in one test, the question that jurors must answer is, "How relevant do you judge this document?" In another study, jurors are asked, "Do you think this document is useful to answer the query?" In addition, multiple questions could be asked for the same document (see Section 2.3 and Fig. 2).

The tool can function with three different types of scales:

1. Free scales allow the juror to enter a number between two defined values. Such a scale can be applied when a study uses wide scales (e.g., percent scales).
2. Closed scales are used when the number of selectable values is relatively low (e.g., a five-point scale to judge the relevancy of a document). Closed scales are shown as menus using radio buttons. When a closed scale is chosen, the number of categories and the category labels can be defined (see Fig. 2).
3. Comment: This scale type allows for open answers and can be used for any observations, such as remarks on the peculiarities of a result.

With flexible questions and scales, we can use RAT to ask any possible question concerning an individual result (or any other type of document represented by a URL). Therefore, RAT can also be used for purposes other than retrieval effectiveness studies, such as annotating test collections or manually classifying documents.

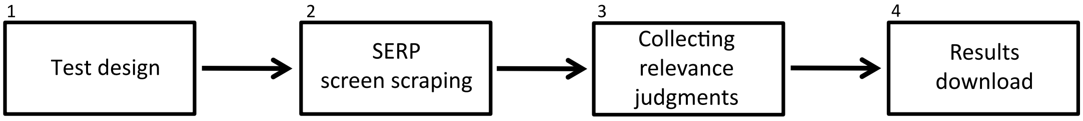
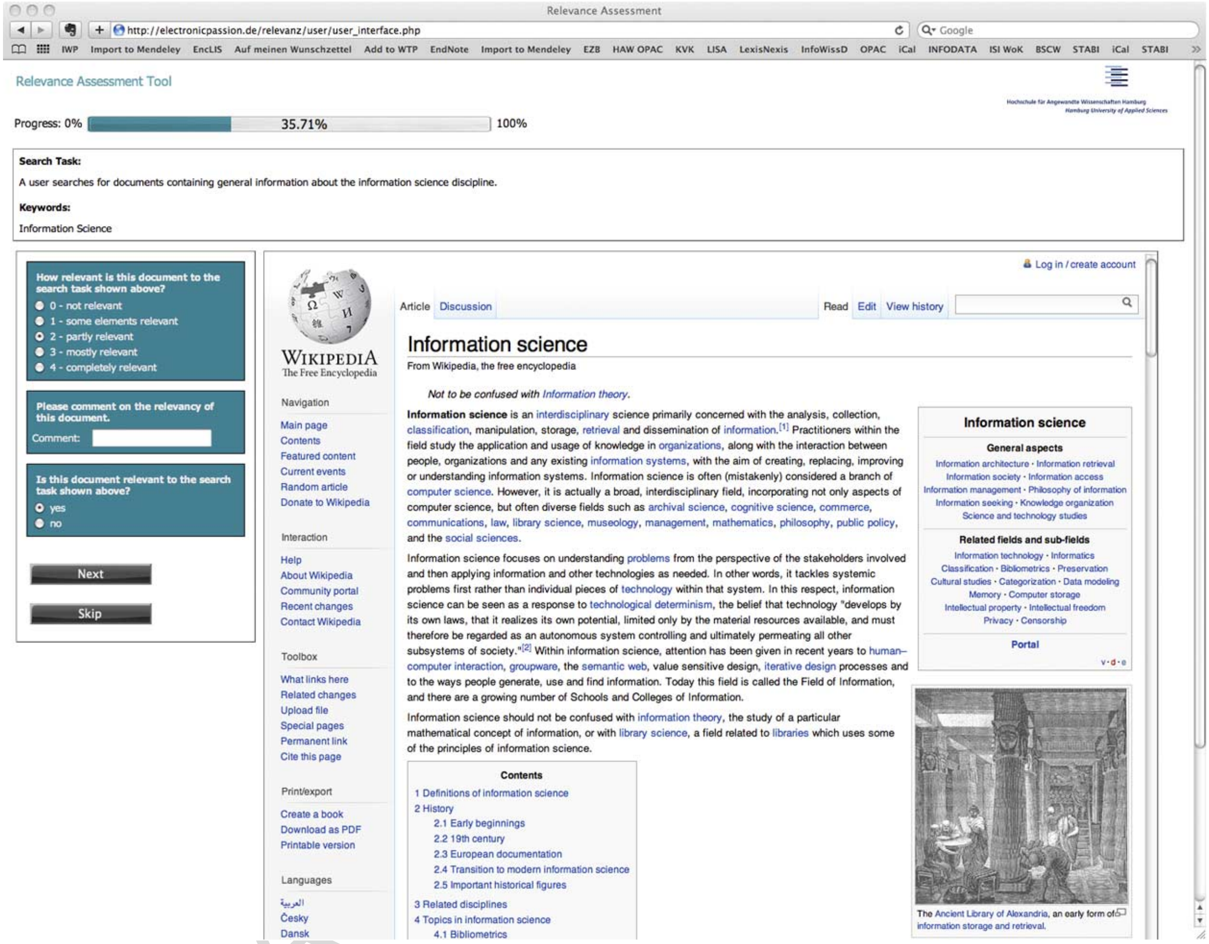


Fig. 1. Process diagram, also showing RAT modules.



Relevance Assessment Tool

Progress: 0% 35.71% 100%

Search Task:

A user searches for documents containing general information about the information science discipline.

Keywords:

Information Science

How relevant is this document to the search task shown above?

○ 0 - not relevant
○ 1 - some elements relevant
● 2 - partly relevant
○ 3 - mostly relevant
○ 4 - completely relevant

Please comment on the relevancy of this document.

Comment:

Is this document relevant to the search task shown above?

● yes
○ no

Next

Skip

Information science

From Wikipedia, the free encyclopedia

Not to be confused with Information theory.

Information science is an interdisciplinary science primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval and dissemination of information.[1] Practitioners within the field study the application and usage of knowledge in organizations, along with the interaction between people, organizations and any existing information systems, with the aim of creating, replacing, improving or understanding information systems. Information science is often (mistakenly) considered a branch of computer science. However, it is actually a broad, interdisciplinary field, incorporating not only aspects of computer science, but often diverse fields such as archival science, cognitive science, commerce, communications, law, library science, museology, management, mathematics, philosophy, public policy, and the social sciences.

Information science focuses on understanding problems from the perspective of the stakeholders involved and then applying information and other technologies as needed. In other words, it tackles systemic problems first rather than individual pieces of technology within that system. In this respect, information science can be seen as a response to technological determinism, the belief that technology "develops by its own laws, that it realizes its own potential, limited only by the material resources available, and must therefore be regarded as an autonomous system controlling and ultimately permeating all other subsystems of society."[2] Within information science, attention has been given in recent years to human–computer interaction, groupware, the semantic web, value sensitive design, iterative design processes and to the ways people generate, use and find information. Today this field is called the Field of Information, and there are a growing number of Schools and Colleges of Information.

Information science should not be confused with information theory, the study of a particular mathematical concept of information, or with library science, a field related to libraries which uses some of the principles of information science.

Contents

1 Definitions of information science
2 History
2.1 Early beginnings
2.2 19th century
2.3 European documentation
2.4 Transition to modern information science
2.5 Important historical figures
3 Related disciplines
4 Topics in information science
4.1 Bibliometrics

Information science

General aspects

Information architecture · Information retrieval
Information society · Information access
Information management · Philosophy of information
Information seeking · Knowledge organization
Science and technology studies

Related fields and sub-fields

Information technology · Informatics
Classification · Bibliometrics · Preservation
Cultural studies · Categorization · Data modeling
Memory · Computer storage
Intellectual property · Intellectual freedom
Privacy · Censorship

Portal

The Ancient Library of Alexandria, an early form of information storage and retrieval.

Fig. 2. RAT user interface. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-130691.)

2.1.3. Language templates

We are based in Europe and plan to allow partners to use RAT to conduct their own studies, so we designed the tool to support studies in languages other than the one in the original user interface. To design a test in another language, a user can simply create a language template, which displays all buttons and system messages in the language chosen. Language templates are reusable throughout projects, so for each language, the template must only be modified once.

2.1.4. Preparing results

A project administrator can choose to have the results for a search task evaluated in the original order or to have the results mixed randomly. Other settings for the preparation of results can be made in the scraper module (see Section 2.2).

2.1.5. User groups

As it is sometimes of interest to have multiple user groups judging the same results, RAT allows administrators to create one or more user groups per test. Jurors access RAT with an access code (see Section 2.3) and are then automatically allocated to their user group. Access codes can be created for certain user groups and for individual users, as well. A project administrator can also choose the number of search tasks a user is allowed to evaluate.

2.2. Search engine results scraping

Some search engines allow automatic querying through application programming interfaces (APIs) (see [16]). However, these interfaces do not produce the same results as the user interface and are therefore not applicable to retrieval effectiveness tests. An alternative is screen-scraping, which extracts results descriptions, results URLs, results types, etc. from the search engine results pages (SERPs). This method allows researchers to collect the same information that a normal search engine user would see [5]. The shortcoming of this method is that every time a search engine changes the HTML code of its results pages, the administrator must adjust the scraper. However, compared to the effort needed to collect the results manually, we think that this method works best for collecting search results data.

We designed our scraper to fetch not only the results themselves, but also the results descriptions, as given by the search engines. Lewandowski [7] showed that the results descriptions are of great importance, as they help the users to decide which results to choose. Unfortunately, these descriptions are often misleading when a seems-to-be relevant description points to a non-relevant result, and vice versa. In the process of scraping the descriptions, the screen-scraper strips all information that could possibly identify an individual search engine (e.g., the phrase "similar results" is unique to Google's results descriptions and could therefore identify this search engine, so the scraper would strip the phrase "similar results" from the descriptions).

The scraper crawls the URLs extracted from the SERPs and saves the contents to a database, because a document's content may change from the moment when the SERP is scraped to the moment when a juror accesses the document through the user interface of RAT. Also, the links within the documents to be judged are replaced with text. They are still clickable in the user interface, but have no function anymore. Instead, a message stating that the link should not be followed in the evaluation process appears.

RAT also supports studies where scrapers are not available for all search engines under investigation. The project administrator can upload an Excel file containing the results descriptions, results URLs, results ranks, and the name of the additional search engine. RAT will then automatically integrate this information into the test. Therefore, it is possible to conduct studies with only scraped results, with no scraper, or a mixture of both.

2.2.1. Scrapers already implemented

We decided to build scrapers for four major Web search engines, Google, Bing, Yahoo and Ask. These are the four scrapers that we continually maintain, while we sometimes build unique scrapers for just one study. For every scraper, we must decide whether building it is worth the effort or if manually collecting the results for a certain search engine would be more efficient, especially if the scraper will only apply to one study. However, we cannot give any general recommendation here, as the HTML code of search engines differs greatly.

Along with our Web search engine scrapers, we built scrapers for shopping search engines (Amazon, Google Products) and music download portals (Musicload, Amazon mp3, iTunes). Because we only used each of these for one study, we decided not to maintain them continually, and will revive them only when we need them for a future study.

2.3. Collecting relevance judgments

Jurors enter the frontend (see Fig. 2) using an access code provided by the project administrator. Users are automatically assigned to the correct project. If the project administrator decided to ask for demographic data, a form is shown to the user. Then, users see a description of the study in which they are taking part, before they start the test.

When a project administrator has decided that both the results and the results descriptions should be judged, the juror either first has to judge all the descriptions before judging the results, or has to judge each description/result pair after the other.

Results are presented within a browser frame. As the documents are saved within the system, we avoid the problem that the documents' content could have changed between the time when the SERPs were scraped and the time when a juror accesses the documents, or even worse, the problem that the document is not available anymore.

A juror working with RAT sees the result document to be judged, a description of the search task, the query, a progress bar, and the questions that judge the document (see Fig. 2).

To motivate jurors to participate in a RAT project, it is also possible to offer rewards in the form of vouchers. An administrator can upload a list of voucher codes (e.g., from Amazon). When a user successfully completes a task, an e-mail with the code is sent. We implemented some basic plausibility check to avoid misuse of the system.

3. Example study: Large-scale retrieval effectiveness test

To show the potential of RAT for efficient search engine retrieval effectiveness tests, we describe a study we did using 1,000 tasks, a very large number compared to the numbers of queries usually used in such evaluations (approx. 50). We took a random sample of queries from transaction log data from a leading German search portal, scraped the results of both Google and Bing, and sent invitations to jurors. Using RAT's crowdsourcing approach, we were able to collect the relevance judgments for the approx. 25,000 results within a few hours. Jurors were given an incentive, namely a gift voucher for Amazon.de worth 5 Euros. Apart from the efficient data collection (both at the results scraping stage, and the relevance judgments collection stage), this approach allowed us for improving the quality of the study, as we were able to not only generate the random sample, but also to have a large amount of relevance judgments collected.

Apart from the study described here, we used RAT for a range of different studies. E.g., in one study, we compared three music download portals, namely Amazon mp3, iTunes and Musicload. We decided to build a scraper for the Musicload and Amazon results. Contrastingly, iTunes is a standalone software and not a Website, so we had to use screenshots from this vendor. We uploaded the screenshots into RAT and combined them with the screen scraping results. Jurors were able to judge the results from all portals within the same user interface.

In another study, we worked together with nutrition scientists at our university. Their aim was to find Websites that sold illegal nutrition supplements. To find relevant pages, they applied the screen-scraping module of RAT to a large number of queries. Results documents were then classified, and the manual classification built the basis for training machine learning algorithms.

4. Conclusion

The Relevance Assessment Tool presented in this paper reduces the time and effort needed to conduct search engine retrieval effectiveness tests. With this tool, researchers can focus more on the scientific work of their studies than on data collection and allocation. The tool's crowdsourcing approach and its easy data collection system allows for more extensive studies, going beyond the normal 10 results per query, the low number of search engines considered, and the limited number of queries. After the introduction of RAT, the main barrier to any study is finding jurors who are willing to participate. At our university, we maintain a proband database for lab studies that we also use for recruiting test participants for retrieval effectiveness studies. However, it soon became clear that we had to offer financial incentives to find enough motivated jurors for larger studies.

Due to the flexible structure of RAT and its modular approach, the tool also allows for studies that go well beyond the primarily intended use case. Also, the modules of the tools can be used independently. When a researcher wishes to compare search engines (or results from other information systems) for which no scraper is implemented, that researcher can upload search tasks and results from the search engines to RAT by using a simple xls file. The researcher can then conduct the process of rating the results and analyzing the data with RAT. Also, RAT can be used as a standalone scraper. Results from search engines can be used for purposes other than retrieval effectiveness studies, such as annotating test collections, classification tasks, or asking jurors to answer multiple questions about the attributes of Web pages.

In future work, we intend to implement a framework for evaluating search engines that includes a "Universal Search" results page design, as described by Lewandowski [10]. With such an extension, researchers could scrape not only the organic results from a SERP, but also additional results, such as advertisements and results from collections beyond the Web (e.g., news, video, blogs). To achieve that, we will combine RAT with a scraping tool we used in another study [5]. Also, we will combine RAT with Search Logger, a tool developed at the University of Tartu (Estonia), which collects user behavior data [14]. We could combine both tools to collect relevance judgments within a process of interactive information retrieval.

As our tool uses a screen scraper component that we must continuously adapt to the current versions of the results pages of the different search engines, we decided against making the tool open source. Versioning would be very difficult, as small changes in the SERPs would require a new version of the tool. However, we are eager to support researchers who wish to conduct search engine tests using RAT.

References

[1] J. Bar-Ilan, The use of web search engines in information science research, in: *Annual Review of Information Science and Technology*, B. Cronin, ed., Information Today, 2004, pp. 231–288.

[2] M. Gordon and P. Pathak, Finding information on the World Wide Web: The retrieval effectiveness of search engines, *Information Processing & Management* **35**(2) (1999), 141–180.

[3] E. Hargittai, The social, political, economic, and cultural dimensions of search engines: An introduction, *Journal of Computer-Mediated Communication* **12**(3) (2007), 769–777.

[4] D. Hawking, N. Craswell, P. Bailey and K. Griffiths, Measuring search engine quality, *Information Retrieval* **4**(1) (2001), 33–59.

[5] N. Höchstötter and D. Lewandowski, What users see – Structures in search engine results pages, *Information Sciences* **179**(12) (2009), 1796–1812.

[6] B.J. Jansen and P.R. Molina, The effectiveness of web search engines for retrieving relevant ecommerce links, *Information Processing and Management* **42** (2006), 1075–1098.

[7] D. Lewandowski, The retrieval effectiveness of web search engines: Considering results descriptions, *Journal of Documentation* **64**(6) (2008), 915–937.

[8] D. Lewandowski, The influence of commercial intent of search results on their perceived relevance, in: *iConference 2011*, ACM, 2011, pp. 452–458.

[9] D. Lewandowski, The retrieval effectiveness of search engines on navigational queries, *ASLIB Proceedings* **61**(4) (2011), 354–363.

[10] D. Lewandowski, A framework for evaluating the retrieval effectiveness of search engines, in: *Next Generation Search Engines: Advanced Models for Information Retrieval*, C. Jouis, ed., IGI Global, Hershey, PA, 2011, pp. 456–479.

[11] D. Lewandowski, New perspectives on search engine research, in: *Web Search Engine Research*, D. Lewandowski, ed., Emerald, Bingley, 2012, pp. 1–16.

[12] D. Lewandowski and N. Höchstötter, Web searching: A quality measurement perspective, in: *Web Search: Multidisciplinary Perspectives*, A. Spink and M. Zimmer, eds, Springer, Berlin, Heidelberg, 2008, pp. 309–340.

[13] M. Machill, M. Beiler and M. Zenker, Search-engine research: A European–American overview and systematization of an interdisciplinary and international research field, *Media, Culture & Society* **30**(5) (2008), 591–608.

[14] G. Singer, U. Norbisrath, E. Vainikko, H. Kikkas and D. Lewandowski, Search-logger – Tool support for exploratory search task studies, in: *SAC2011*, ACM, 2011.

[15] J. Tague-Sutcliffe, The pragmatics of information retrieval experimentation, revisited, *Information Processing & Management* **28**(4) (1992), 467–490.

[16] F. Tosques and P. Mayr, Programmierschnittstellen der kommerziellen Suchmaschinen, in: *Handbuch Internet-Suchmaschinen*, D. Lewandowski, ed., Akademische Verlagsgesellschaft Aka GmbH, Heidelberg, 2009, pp. 116–147.

[17] J. Véronis, A comparative study of six search engines, 2006, available at: http://sites.univ-provence.fr/veronis/pdf/2006-comparative-study.pdf.

[18] M. Zimmer, Web search studies: Multidisciplinary perspectives on Web search engines, in: *International Handbook of Internet Research*, J. Hunsinger, L. Klastrup and M. Allen, eds, Springer, Dordrecht, 2010, pp. 507–521.