Relevance and utility in an educational search environment

Theo Huibers*
University of Twente
Enschede, The Netherlands
t.w.c.huibers@utwente.nl

Thijs Westerveld
Wizenoze
Amsterdam, The Netherlands
thijs@wizenoze.com

*Huibers is co-founder of Wizenoze

ABSTRACT

Technology is increasingly being used in education. Children are also increasingly using search engines when searching for information on certain themes. The question of what is a good search system to use in education has still not been answered definitively. In this article we explain the steps Wizenoze takes to build and evaluate a good search system. We split our analysis in the way Cooper already proposed in 1971 as the two aspects of relevance: logical relevance and utility.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Social and professional topics** → **Children**; • **Information systems** → *Evaluation of retrieval results*; Information retrieval; • **Human-centered computing** → *HCI design and evaluation methods*; Human computer interaction (HCI).

KEYWORDS

Children, information retrieval, evaluation, education

ACM Reference Format:
Theo Huibers and Thijs Westerveld. 2019. Relevance and utility in an educational search environment. In *KidRec '19: Workshop in International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems, Co-located with ACM IDC, June 15, 2019, Boise, ID*. ACM, New York, NY, USA, 6 pages.

KidRec '19, June 15, 2019, Boise, ID

INTRODUCTION

Technology is rapidly making its entry into education, from serious games [6] to robots [2], technology changes the classroom and redefines the way of learning. One of the most important learning tasks for children is to develop their knowledge and learn about specific themes at their own reading level e.g. in the field of history, geography, and social studies. Exploration of corresponding themes according to their learning lines is the foundation of the current educational system.

Technology in general and Information Retrieval Systems (**IRS**[1]) in particular can play an important role in the exploration of information. In addition to the classic educational books and libraries, children today spend considerable time online, searching and receiving information from various websites and apps. While searching for educational information, school children use search systems to locate resources and receive site recommendations that might be useful for them. Some sites and apps are specifically designed for children, others are intended for adults, but widely used by children including Google, the most popular search site.

The call for good, reliable, child-friendly systems especially in an educational environment has been made many times and the thesis that the algorithms and interfaces of "adult" information systems are not necessarily suitable or fair for children is widely accepted [8]. For example, searching for a topic such as 'Volcano' with Google as an eight-year student brings him or her to restaurants, complex Wikipedia pages, music bands, and all kinds of content that is too difficult for the learner to understand due to their reading ability.

However, there is still no clear and balanced view on what makes a good search system for children, nor on what content should be considered good enough to be retrieved or recommended. In our opinion an educational information retrieval system for children can be described as good when it returns information that is readable, relevant and reliable for the child.

The structure of the article will be as follows. First we will explain the context and evaluation of IRS. Next we will present Wizenoze and its educational search system called Web for Classrooms. We will go onto explain how Wizenoze analyses the 'goodness' of their system by inspecting the results on relevance and utility before finally presenting our future research.

EVALUATING INFORMATION RETRIEVAL SYSTEMS IN GENERAL

Information Retrieval Systems (**IRS**) are software tools that provide diverse users with resources that are **relevant** to their corresponding information needs [4]. Relevance is a fundamental concept for information retrieval [11]. In 1971, Cooper [3] distinguishes two aspects of the notion of relevance:

- Logical Relevance, which describes whether a retrieved document has some topical bearing on the information need in question, and
- Utility, which describes the ultimate usefulness of the retrieved document.

[1]We use the general term information retrieval system to describe a whole set of systems selecting and ordering information given a certain information need such as web search systems, recommender systems and filtering systems

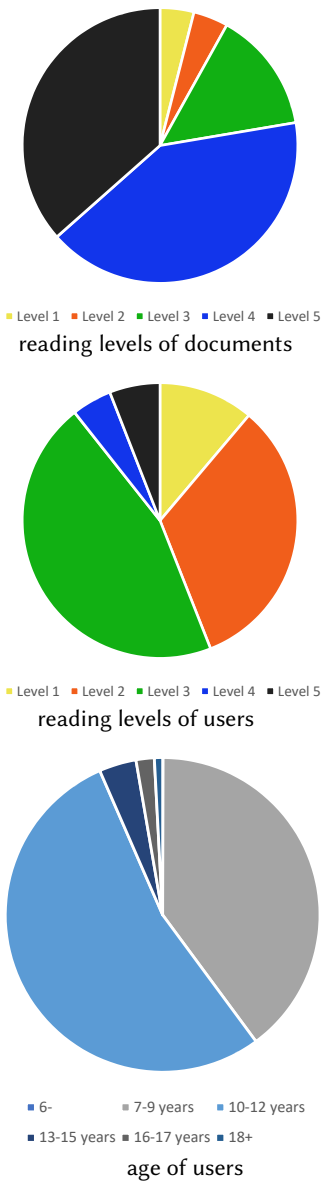

Figure 1: Document and user statistics for Web for Classrooms

Large-scale evaluation and comparison of algorithms that define the notion of (logical) relevance started in 1992 at the annual Text REtrieval Conference (TREC). In November 1992 the first TREC was held. Its proceedings contain papers about search experiments and their results. Benchmarks are the key elements of the TREC program, an officially organised activity, which has as its main goal to study different approaches to the retrieval of text for large and specific document collections. More than twenty-five years later, TREC still is the major experimental effort in the information retrieval field [12]. To compare the results obtained there is a detailed schedule with specific information tasks that all the participants of TREC should obey.

Utility is assessed in TREC as well, but not on the same scale. There has been an *Interactive Track* at TREC that looked at IRS in context. The same topic is central to the CHIIR conferences that started in 2016 [1].

IN THE CONTEXT OF IRS FOR CHILDREN

While IRS for adults have been studied and evaluated for several decades in all its diversity, e.g. the yearly organised TREC, the evaluation of IRS for children has hardly been studied. A number of studies have looked into children's information seeking behavior [5, 7, 9, 10], but we know of no effort to define the quality criteria like relevance and utility for such systems, and how to measure those.

WIZENOZE

Founded in 2013, Wizenoze aims to offer students a closed information domain with educational information gathered from the internet in combination with a child-friendly search engine, Web for Classrooms (**WfC**). The search engine delivers information that is readable, relevant and reliable.

WEB FOR CLASSROOMS

Web for Classrooms helps students find relevant content online. We employ a mix of human and machine intelligence to bring the chaos of the web down to a ranking of readable and reliable search results. The first step in the process is to identify sources (websites) that need to be included in our collection. This step is largely manual: we assess the reliability of a source and its suitability for an educational context. Next, our crawlers visit the source to collect individual web pages from it and keep them up to date in our collection.

The readability of each document in the Web for Classrooms collection gets classified on a five point reading scale. The classifier is trained on a proprietary collection of labeled texts from a heterogeneous set of sources including news, web data and textbook material. We compute a wide range of textual features ranging from low level text statistics (e.g., average word length, average sentence length)

Participating schools

- Burntwood School (Secondary)
- Chestnut Grove Academy (Secondary)
- Hampstead School (Secondary)
- Graveney School (Secondary)
- Gesher School (Primary/SEN)
- Grove Road (Primary)

Curriculum areas covered

- KS1, 2, 3 and 4
- SEN, Literacy, Music, Science, English and History

Sidebar 1: Details of the schools and topics included in the classroom intervention study

to more advanced features including vocabulary use, noun variation, type token ratio, and average number of passives per sentence.

Finally, all documents are indexed into an Elasticsearch cluster[2]. For search, a request consists of the search terms as well as an indicator of the users' age and reading level. In the ranking of the results, we combine topical features ($tf.idf$ based text matching on various fields), and document quality features like readability, suitability and recency.

The Web for Classrooms is currently available for Dutch and English content. It holds over 12 million documents in total and is expanding every day. In the Netherlands, we now reach 78% of all schools only 2 years after launch. In the UK we are rapidly approaching a similar percentage through partnerships with educational platforms and educational publishers. Figure 1 shows general statistics about the Web for Classrooms collection and its user base (Dutch and English combined).

RELEVANCE AND UTILITY FOR WEB FOR CLASSROOMS

To measure the quality of the results that we present to students in Web for Classrooms, we look at both (logical) relevance and utility. For *relevance*, we develop test collections following the IRS evaluation paradigm as described above. For *utility*, we are interested in the actual use of the results. As an indirect metric of this, we monitor user engagement. Returning, active and engaged users clearly found some value in the system; we can therefore assume they found utility in the results. To monitor utility more directly, we perform classroom intervention studies and measure learning outcomes with and without the support of Web for Classrooms.

Test collection

A good IRS test collection mimics reality as closely as possible. This means documents, queries and relevance judgements need to be representative of the real user situation.

For documents this is straightforward: a representative subset of the Web for Classrooms collection is sufficient. Realistic queries and relevance judgments are a bit harder to get with children or students as the main target audience. Ideally, we would have students formulate representative search queries and explicitly assess the relevance of the search results for these queries. However, we find that students, especially younger ones, have trouble assessing the relevance of a result. They typically overestimate their own skills in reading and understanding a text, and they tend to give socially desirable answers.

In our test collection, both factors would lead to false positives: irrelevant documents that are judged as relevant.

To some extent these issues can be avoided by monitoring student behaviour in practise, using result clicks as implicit judgements, but for our test collection, we employ adults to mimic younger searchers. We ask them explicitly to think of their information need imagine as a topic for a student's

[2]https://www.elastic.co/

ASSESSOR INSTRUCTIONS

What is relevant?

- The starting point for judging is the detailed description of the information need.
- Think from the student's perspective. As a rule of thumb, you can imagine doing a project assignment on the topic. Would you use this information?
- Don't think from the search engine's perspective: Thinking, "I understand why it would return this result", does not make it a good result or relevant for an end user.

Duplicate information.

- Judge items in isolation
- If you see duplicate documents, almost duplicate documents, or just the same relevant information in two different documents, please mark all occurrences as relevant.

Age and reading level.

- We do not take age or reading level into account in the test collection. In this way, we are trying to isolate the quality of the ranking algorithm.
- Do not mark items as irrelevant results just because they are too hard, or not suitable for a specific age group.

Title, image and snippet quality.

- The test collection is for assessing the ranking quality, not the quality of the title, image, or snippet.
- Judge the relevance of the document; don't mark items as irrelevant if for example only the image is missing.

Sidebar 2: The instructions we gave to assessors for judging relevance.

project assignment on the topic. The detailed instructions we give our assessors are shown in Sidebar 2. To make sure that the topics in the test collection are close enough to real topics, we take them from the school curriculum and from our search logs. The documents in a test collection are typically kept static, while real life collections are expanding and changing every day. We choose to run our evaluation on the live Web for Classrooms collection, and make sure we keep reassessing documents as the rankings change. In this way, we can use the same collection as a quality assurance tool to monitor the live results for important queries.

Classroom intervention studies

To get more insight into how Web for Classrooms is used in practice, we perform regular school visits. We conduct teacher and student interviews to gain insight into the utility that Web for Classrooms brings. Moreover, we have performed a classroom intervention study that we outline below.

Working with the London Grid for Learning over a three-month period in 2018, Wizenoze conducted research in six London schools. The six institutions provided us with access to different key stages and curriculum areas so that insights on the educational value of WfC could be collected (see Sidebar 1 on page 4 for details). We explored WfC as a whole institution proposition regardless of key stage or curriculum area.

In the study, an a/b testing approach was applied. Teachers were asked to continue with their programmes of study, but when a research task was been asked, half their students (group a) used Web for Classrooms and the other half (group b) used their usual search tool (for example Google) to support their task. On completion, the teachers sent us anonymous samples of work from both the a and b group. Within each sample batch (class) students were labelled as exceeding expectations, as expected, and below expectations. A comparative analysis between the groups was then carried out. The results indicated that WfC has a clear impact on improving learning outcomes with

- 91% of sampled work showed students using the Web for Classrooms progressed further towards the desired learning outcomes;
- 18% of sampled work completed less of the work expected in the time given,even though they had progressed further towards the desired learning outcomes;
- 82% of sampled work gave responses with more factual information to support the desired learning outcomes;
- 73% of sampled work gave more examples to illustrate understanding.

The intervention study shows that Web for Classrooms helps students improve their learning outcomes. Some students took longer to complete a task compared to the control group even though their answers were often better. They do take some more time. Further analysis is needed to see if this is due to unfamiliarity or whether for example a revision of the user interface is required.

Future research

Now that we have an evaluation framework for Web for Classrooms, we can further scale up our evaluations. We can investigate the differences in user search behavior and at the same time study the quality of educational search systems. We can analyze the impact of reading level and age on search quality, and we can investigate cultural differences in search effectiveness.

REFERENCES

[1] 2019. *CHIIR '19: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA.

[2] Fabiane Barreto Vavassori Benitti. 2012. Exploring the educational potential of robotics in schools: A systematic review. *Computers Education* 58, 3 (2012), 978 – 988.

[3] W.S. Cooper. 1971. A definition of relevance for information retrieval. *Information Storage and Retrieval* 7, 1 (1971), 19 – 37.

[4] W Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Search engines: Information retrieval in practice*. Vol. 283. Addison-Wesley Reading.

[5] Tatiana Gossen. 2016. *Search engines for children: search user interfaces and information-seeking behaviour*. Springer.

[6] Frederik De Grove, Jeroen Bourgonjon, and Jan Van Looy. 2012. Digital games in the classroom? A contextual approach to teachers' adoption intention of digital games in formal education. *Computers in Human Behavior* 28, 6 (2012), 2023 – 2033.

[7] Hanna Jochmann-Mannak, Theo Huibers, Leo Lentz, and Ted Sanders. 2010. Children searching information on the Internet: Performance on children's interfaces compared to Google. In *SIGIR*, Vol. 10. 27–35.

[8] Hanna Jochmann-Mannak, Theo Huibers, and Ted Sanders. 2008. Children's Information Retrieval: Beyond Examining Search Strategies and Interfaces. In *Proceedings of the 2Nd BCS IRSG Conference on Future Directions in Information Access (FDIA'08)*. BCS Learning & Development Ltd., Swindon, UK, 8–8.

[9] Hanna Jochmann-Mannak, Leo Lentz, Theo Huibers, and Ted Sanders. 2012. Three Types of Childrenś Informational Web Sites: An Inventory of Design Conventions. *Technical communication* 59, 4 (2012), 302–323.

[10] Nikos Manouselis, Hendrik Drachsler, Riina Vuorikari, Hans Hummel, and Rob Koper. 2011. Recommender systems in technology enhanced learning. In *Recommender Systems Handbook*. Springer, 387–415.

[11] Stefano Mizzaro. 1998. Relevance: The whole history. *Journal of the American Society for Information Science* 48, 9 (1998), 810–832.

[12] Ellen M. Voorhees and Angela Ellis (Eds.). 2018. *Proceedings of The Twenty-Seventh Text REtrieval Conference, TREC, Gaithersburg, Maryland, USA*. Vol. Special Publication 500-331. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec27/trec2018.html