Finding Quality: A Multilingual Search Engine for Educational Research

Aaron Kaplan, Ágnes Sándor, Thomas Severiens, Angela Vorndran

Short Summary
To develop a field specific and multilingual search-engine, numerous algorithms are needed in addition to a general-purpose search engine. Here we describe the focal areas of development done in EERQI: Automatic classification for educational research, multilingual retrieval, query extension and relevance ranking. The classification algorithms, developed in EERQI enable a crawler to identify relevant objects with respect to a scientific field; the multilingual algorithms allow the retrieval of documents in several languages; query extension proposes related query terms to the user; relevance ranking is enhanced by semantic analysis.

1 An Automated Decider: Which Objects are Relevant for Educational Research?

Having a general web search engine, it would be impossible to decide which of the harvested objects are relevant for Educational Research, and which ones are not. One could only select the starting addresses for the crawling process wisely, but it would be impossible to detect new clusters of relevant material online in an automated way. To avoid this constraint, we developed and tested an algorithm deciding which of all crawled objects may be of relevance for Educational Research.

To train this machine-based learning algorithm, it was necessary to extract a number of full texts from the EERQI database of published articles and books. As the developed algorithm is highly sensitive to the language of the object to be tested, we had to train four different algorithms for the four EERQI languages: English, French, German, and Swedish. At least for the German and English algorithms, we had a sufficient number of training objects.

The technique used for the algorithms is quite old and well tested, but before the age of Cloud Computing, it was hard to find use-cases small enough to be implemented in real scenarios. Thus, one of the challenges was to boost the technical implementation and to make it usable.

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_2, © Springer Fachmedien Wiesbaden 2014

The technology used for duplicate detection is described by e.g. Monika Henzinger (2006). Her work is based on algorithms developed by Broder in 1995-1997, who in turn refined algorithms described theoretically by Rabin (1981). The technology described by Henzinger is current state of the art for comparing big textual collections. Sorokina et.al. (2006) describe some basic rules, to reduce the number of shingles (an *k* words long phrase is called a *k-shingle*) to be handled, such as the rule to remove all shingles crossing sentence boarders, to remove capitalization, to replace stop words by an asterisk, etc. Empirical tests showed that 4-shingles are the optimum size for our deciding algorithm.

We made use of all these rules and trained deciding algorithms for all the four EERQI languages, using published articles and books as in-put. We were careful to train the algorithms, taking into account information on authors, publishers, from any genres and subfields in Educational Research. As the number of available publications in Swedish was too low, we decided to focus our activity on English, French and German. For the French algorithm, we had to add several articles from other sources to reach the necessary amount of documents for training, which is about 500 full texts. At the end, the tests showed that only the German and the English algorithms were usable, while the other two were unable to appropriately take into account information on subfields.

Part of the training procedure is to have a 'negative group' of full texts from other, but ideally adjacent fields, where phrases (shingles) available in both text collections are removed from the list of field specific phrases. At the end, one has a list of uni-lingual phrases (shingles) which are typical for Educational Research and represent the whole field. Most programmers call this kind of list a ’finger print’.

We made use of these finger prints to compare them with the list of shingles extracted from objects to be tested. If the percentage of shingles extracted from the object, and also being available in the finger print, exceeded a critical value (individually determined for every finger print case), an object was marked as being of potential relevance for Educational Research.

This service was coupled to the search engine using a REST-based[5] webservice. This allows other software to connect to our service in a defined and open way.

To test our algorithms, we used 50 relevant and 50 non relevant documents from the EERQI database and from other Open Access institutional repositories. Out of the relevant documents, the algorithm for English documents identified 91% as relevant, while 3% of the relevant documents were not identified. The

[5] REST: “Representational State Transfer”, a dialect for a web service

corresponding results when testing the German algorithm was a recognition rate of 89% of the relevant documents while missing 5%, i.e. results that were slightly worse; and the French algorithm only recognized 73% while failing to identify 12% of the relevant documents. We could not develop and test a Swedish algorithm because there are too few publications available for training and testing.

The developed software, as well as all fingerprints are published under the BSD[6]-license on the EERQI web-server[7], to be reused by other projects. It already has been re-used in the field of biotechnology[8].

2 Multilinguality and query expansion

To enhance the field-specific search engine we built a software module that performs query translation and identifies relevant term suggestions, and we created a user interface that makes this functionality available to users via the web.

To support query translation and term suggestion, we use a number of different lexical resources: term networks that were compiled by DIPF and IRDP expressly for the purposes of this project, existing multilingual controlled vocabularies (TESE9, EET, and TheSoz10), and the general-purpose (i.e. not education-specific) query translation service from the CACAO project11 (the CACAO query translation service was graciously provided to the EERQI project by CELI). To translate a query, the software tries first the term networks, then the controlled vocabularies, and finally the CACAO service. For term suggestion, only the term networks and the controlled vocabularies are used. We also integrated into the query translation and suggestion software the same linguistic processing modules that were used in the indexer, so that the base forms of query words can be matched with the base forms of words in the indexed documents.

We built a web interface that allows the user to enter a query in English, French, German, or Swedish, and retrieve documents in one or more of these languages. Results for all desired languages are returned in a single list, ranked by estimated relevance to the query. When term suggestions are available, they are displayed (in the query language) next to the results. Clicking on a suggestion causes that term to be added to the query. In an earlier version of the interface we allowed the user to modify how the query was translated, but testing

[6] BSD-license: An open source license, formerly known as Berkeley Software Distribution

[7] Decider Software and Fingerprints published at: http://www.eerqi.eu/sites/default/files/EERQI-Classifier-and-Fingerprints.tar

[8] http://www.bibliometrie.info/forschung/teilprojekte.html

[9] http://ec.europa.eu/education/news/news1907_en.htm

[10] http://www.gesis.org/en/services/tools-standards/social-science-thesaurus/

[11] http://www.cacaoproject.eu/

indicated that some users were confused by this functionality, so in the current version the translation is displayed but cannot be modified. Users who are not satisfied with the automatic translation can simply use a monolingual search in the target language.

To determine how well the multilingual search functionality works and to identify opportunities for improvements, we performed several rounds of user testing of increasing size and formality. The initial rounds involved a few participants among the EERQI partners. After taking into account the feedback from the earlier rounds, we ran a larger set of tests in which education researchers worldwide were invited to participate.

While there are some testing methodologies for comparing cross-language information retrieval systems that have emerged as standards in the research community, these techniques are only applicable when the systems being compared are used to index the same set of documents, and when the query process consists merely of submitting a textual query and retrieving a list of results. Since the EERQI content base was compiled expressly for this project, it has not yet been indexed by any competing search engine; and since our search engine allows interactive query refinement via term suggestions, an evaluation methodology designed for one-shot query mechanisms is not applicable. In light of this, our goal in designing a testing methodology was not to compare our system directly to others, but to identify opportunities for improvement and to establish tools for tracking improvements from one version of our system to the next.

A number of independent factors affect the quality of search results, including coverage and quality of the collection being searched, of the lexical resources used, of the linguistic software for finding base forms, the appropriateness of the ranking formula, and the design of the user interface. To have a detailed understanding of the performance of the system, it would be interesting to design tests that isolate each of these factors. In some cases this would also facilitate comparison with other search engines. However, given the resources allocated, such detailed evaluation was out of the scope of the EERQI project. In some cases subsystems have already been evaluated elsewhere, e.g. the CACAO query translation system has participated in the CLEF evaluation campaign (Bosca and Dini 2009).

The evaluation methodology has two parts: quantitative analyses of user log data, and qualitative feedback in the form of a questionnaire and interviews.

Quantitative measurement:

Each time a query is submitted, the server logs an entry that includes a timestamp, the text of the query, the method that was used to submit the query (typing in the query box or clicking on a term suggestions), the query language and the requested result languages, and any term suggestions made by the system. When a user clicks on a link in the result list to read a document, or advances to a subsequent page of results, these clicks are also logged and associated with the query from which the result list was generated.

In the first two weeks of the final round of testing, 1152 queries were logged in 289 sessions, where a session corresponds (roughly) to a series of queries made from the same computer within a period of ten hours. 46% of the queries submitted were cross-language searches. The total number of documents viewed was 516, or 0.45 documents per query on average. More specifically, in 81% of the cases, none of the results were viewed; in 10% of the cases one document was viewed; in 4% of the cases two documents were viewed; and in the remaining 5% of the cases three or more documents were viewed.

One measure of the quality of a query translation system is the ratio of cross-language search performance to monolingual search performance. With an ideal query translation system, one would find as many relevant results by using automatic translation as one does when searching in each language separately, resulting in a ratio of cross-language performance to monolingual performance of 1. In our tests, the average number of viewed documents per query was .30 for cross-language queries and .57 for monolingual queries, for a ratio of .53.

The system suggested additional terms for 81% of the queries. In cases where suggestions were made, the user clicked a suggestion 12% of the time.

Qualitative feedback:

All test participants were requested to fill out a questionnaire after using the system, but we made no attempt to enforce compliance with this request. We received 15 questionnaire responses, which is only 5% of the number of sessions observed on the search engine. Reactions were generally quite positive, but since the respondents were self-selected and the response rate was so low, statistics compiled from the responses would be difficult to interpret. The value of the responses is primarily that they describe problems that users encountered, indicating ways in which we can improve the search engine in the future.

In addition to the questionnaire, which was widely distributed via email lists, we contacted a small number of users personally to arrange telephone interviews to discuss their experiences in depth. We have performed five such interviews.

The most frequent comments in the questionnaire responses and the interviews were the following:

Many users requested an ”advanced search“ mode that gives more control over the search, particularly Boolean operators and constraints on metadata fields, e.g. constraining the search to documents published in certain years.

This remark was often linked to the complaint that a search returned ”too many results“, leaving the users with a need for options to cull the list. Since results are ranked according to a scoring function giving higher scores to documents with more query terms, a document containing all query terms would be at the top of the list. Our expectation was that users would be reading the list from the top down, and then stops reading when they perceived that the remaining results were no longer relevant. However, feedback shows that many users read the whole list without considering any difference in relevance of the retrieved documents.

This mismatch in expectations is related to the difference between curated electronic library catalogs and web search. Curated collections typically have rich and reliable metadata, and support Boolean search with field constraints, whereas web search engines rely on ranking-based techniques with less user intervention in order to deal with noisier, non-curated data. Since the EERQI document base is a mixture of curated data from publishers and non-curated documents from the web, we chose to use a web-style approach, but testing revealed that many users were expecting a tool similar to a digital library. If we have an opportunity to develop the system further, we will approach this problem in two ways: by making a more fine-grained control of the search terms when possible; and by better managing user expectations, e.g. by explaining the ranking criteria.

Several users complained that the "title" metadata field was often missing or containing inadequate or irrelevant information. This is, again, a result of using documents crawled from the web, with metadata extracted by an error-prone automatic method rather than curated. It will never be possible to achieve 100% accuracy in automatically-extracted metadata, but there may be ways to improve on the methods we are currently using.

Translation of German compound words was often seen to be problematic. When a German compound word is not present in the term networks, its individual components are translated independently, and documents containing the translations are retrieved. This proved to be too broad in many cases. To narrow the search, it might be preferable to set as requirement that the individual components of translated compound words occur near each other.

3 Enhancing Relevance Ranking

In Chapter 4 we outlined a method for defining and detecting salient sentences in social science research articles. Similarly to the way content-oriented metadata – title and abstract - are used in digital libraries, we have used these sentences as additional metadata in the EERQI search engine, and we tested the performance.

The basic algorithm applied by the search engine includes term frequencies (TF) and inverse document frequencies (IDF) for ranking the retrieved documents. These measures are based on the frequency of occurrence of search terms in the documents. The so-called TF-IDF formula weighs the number of times a search term occurs in a document against the number of times a term occurs in the whole document collection. If a search term thus appears in one document frequently but only rarely or not at all in most of the other documents in the document collection, the document is ranked highly (cf. Manning et al., 2009).

The method developed for the EERQI search and query engine is meant to support the ranking of retrieved documents by assigning a higher weight to the query terms retrieved in sentences detected as salient sentences by XIP (see Chapter 5). We suggest that as a consequence the precision concerning the relevance of the retrieved documents will increase, since the likelihood that the query term represents the content of the whole document rises. While a retrieved term with the TF-IDF method can be located in any part of the document and thus may be irrelevant to the gist and main content of the article, a term retrieved in a salient sentence bears high resemblance to the general topic of the article.

In the following paragraphs we provide indications for comparing the results provided by basic EERQI search engine with the query “sport AND school”. We evaluated[12] an article as relevant if its main topic was related to both school and sport.

We evaluated the relevance of the first 15 articles returned by the basic relevance ranking algorithm. Our evaluation found 3 relevant articles with respect to the query. None of these articles were selected as relevant by XIP.

XIP selects an article as relevant with respect to the query if it contains at least one salient sentence that contains both query words. We evaluated our tool on the 330 articles (out of the 1200 retrieved by the basic search engine) that contain at least one sentence with both query words.

Out of the 330 articles 85 were selected by our program, i.e. in 85 articles at least one salient sentence contained both query words.
The following list shows the human evaluation of these 85 articles:

- The number of relevant articles according to human evaluation: 23 (most of these are ranked low by Lucene)
- In 4 articles out of these the salient sentence is detected on an erroneously selected sentence
- The number of non-relevant articles according to human evaluation: 62

[12] The evaluation was carried out independently by the two authors. The inter-annotator agreement was almost 100%.

Analysis of the errors:

- Error due to format transformation[13]: 29
- The automatic sentence-type detection is correct but the sentence is not relevant with respect to the query: 15
- The automatic sentence-type detection is correct and the sentence is relevant with respect to the query, but the whole article is not relevant: 7
- Erroneous sentence-type detection: 11

Out of the remaining 245 articles, 35 have been evaluated as being relevant to the query. In these articles, we checked sentences containing both query words to search for salient messages that were missed by the tool, and we found one such example.

In all we found 58 relevant articles while evaluating our tool. They were all ranked low (beyond 100) by the basic ranking algorithm.

This test allowed us to conclude that salient sentences detected by XIP are indicators of relevance for queries, and they provide complementary results with respect to the TF-IDF method. Salient sentences have been given additional weight in the final EERQI search engine, and they are also used as snippets that present the retrieved documents.

4 References

Bosca A., L. Dini, Cacao Project at the TEL@CLEF Track. Working Notes for the CLEF 2009 Workshop, Corfu, Greece. ISSN: 1818-8044

Henzinger, Monika (2006): Finding near-duplicate web pages: a large-scale evaluation of algorithms. SIGIR '06 Proceedings. New York: ACM

Manning, C.D., Raghavan, P. & Schütze, H., 2009. *Introduction to Information Retrieval.* Online edition. Cambridge: Cambridge University Press.

Rabin, M. (1981): “Fingerprinting by random polynomials”. Report TR-15 81, Center for Research in Computing Technology, Harvard University.

Sándor, Á., Vorndran, A. (2010): Extracting relevant messages from social science research papers for improving relevance of retrieval. Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires, Argentina, 10-14 May 2010.

[13] The retrieved articles (pdf, html, doc) are transformed to plain text for the NLP analysis

Sorokina, Daria; Gehrke, Johannes; Warner, Simeon; Ginsparg, Paul (2006): "Plagiarism Detection in arXiv". http://www.computer.org/plugins/dl/pdf/proceedings/icdm/2006/2701/00/270101070.pdf

Springer

http://www.springer.com/978-3-658-05968-2

Assessing Quality in European Educational Research
Indicators and Approaches
Gogolin, I.; Aström, F.; Hansen, A. (Eds.)
2014, XIV, 201 p. 27 illus., Softcover
ISBN: 978-3-658-05968-2