A NEW APPROACH TOWARDS VERTICAL SEARCH ENGINES
Intelligent Focused Crawling and Multilingual Semantic Techniques

Sybille Peters, Claus-Peter Rückemann, Wolfgang Sander-Beuermann
Regionales Rechenzentrum für Niedersachsen (RRZN), Leibniz Universität Hannover (LUH), Hannover, Germany
peters@rrzn.uni-hannover.de, rueckemann@rrzn.uni-hannover.de, wsb@rrzn.uni-hannover.de

Keywords: focused crawling; search engine; vertical search engine; metadata; educational research; link analysis.

Abstract: Search engines typically consist of a crawler which traverses the web retrieving documents and a search front-end which provides the user interface to the acquired information. Focused crawlers refine the crawler by intelligently directing it to predefined topic areas. The evolution of search engines today is expedited by supplying more search capabilities such as a search for metadata as well as search within the content text. Semantic web standards have supplied methods for augmenting webpages with metadata. Machine learning techniques are used where necessary to gather more metadata from unstructured webpages. This paper analyzes the effectiveness of techniques for vertical search engines with respect to focused crawling and metadata integration exemplarily in the field of "educational research". A search engine for these purposes implemented within the EERQI project is described and tested. The enhancement of focused crawling with the use of link analysis and anchor text classification is implemented and verified. A new heuristic score calculation formula has been developed for focusing the crawler. Full-texts and metadata from various multilingual sources are collected and combined into a common format.

1 INTRODUCTION

This investigation is part of a an ambitious scheme funded by the European Commission under the 7th Framework Programme: The European Educational Research Quality Indicators (EERQI) project was launched 2008 for a duration of three years with the purpose of finding new indicators and methodologies for determining research quality of scientific publications in the field of "educational research" (EERQI-Annex1, 2008). A key task within this project is the development of an exemplary vertical search engine for "educational research" documents. For this purpose, mechanisms must be found for locating "educational research" publications in the WWW as well as for distinguishing scientific research documents from non-scientific documents.

The goal is to provide extensive search capabilities for the user of the search engine. It should be possible to search within the full-text and metadata (language, publisher, publication date, peer-review status etc.) of the document. The methods described exemplarily in this paper might then be applied to any vertical search engine

1.1 Existing Search Engines

A number of search engines focusing on scientific research were analyzed. These included OAIster (OAIster, 2009), Scirus (Scirus, 2009), Google scholar (Google Scholar, 2009), the Education Resources Information Center (ERIC) (ERIC, 2009) and the British Education Index (BEI) (BEI, 2009). None of these search engines provided all required features including a granular topical selection, searching within content and metadata such as title, subject, author and/or language and inclusion of non-english documents in the corpus.

1.2 Focused Crawling

The goal of a focused crawler is to limit crawling to a specific field of interest. Frequency measures for the keywords within URLs (Zheng et al., 2008), link anchor texts (Zhuang et al., 2005), title and full-text (Bergmark et al., 2002) as well as occurrences of links to and from other pages (Chakrabarti et al., 1998) are some of the parameters that have been evaluated. Machine learning methods have been applied to steer the crawler to pages with a higher probability of compliance with the requested topic and determine whether the documents meet the criteria. The crawl must also be refined by selecting a list of high quality start URLs ("seeds"). Starting the crawler on open access archives, academic author or institutional websites are some of the strategies that have been used (Zhuang et al., 2005). It must also be established how far the crawler may deviate from topically relevant pages ("tunneling") to find more clusters of relevant pages (Bergmark et al., 2002).

Classification methods are then applied to determine whether the retrieved page belongs to a target topic. For this purpose a set of training documents are used. Common methods are a vector space model with tf-idf term weights and a Naive Bayes, k–nearest neighbor or support vector machine classifier (Pant et al., 2004; Manning et al., 2008). In addition to limiting the crawl to a certain topic, mechanisms must be applied to assure that the retrieved documents are actually scientific research documents. It needs to be determined how well existing classifiers are capable of doing this. This remains unclear in some of the existing publications. For example, Zhuang et al. (Zhuang et al., 2005) have measured the recall level of their approach but not the precision. In order to retrieve only academic documents, the documents themselves may be analyzed for structure or content as well as the sites on which they are located.

1.3 Metadata

Metadata in this paper means any information further describing a digital "educational research" document (referred to as ERD in this paper). This may be bibliographic information such as the title, author, publisher, abstract or ISSN which are assigned by the time of publishing and additional keywords or quality criteria which may be automatically or manually attributed to the document. It may also be the language, file format or number of pages.

Metadata is useful for searching and browsing within a set of documents (Witten et al., 2004). Combining full-text search with metadata search greatly enhances the capabilities of the user to refine queries. Displaying the metadata in the search results provides additional valuable information. Sources for gathering metadata may be combined. Metadata may also be extracted from the full-text documents themselves. Extensive research has been done on various methods to achieve this, for example using a support vector machine based classification method (Han et al., 2003).

The Web pages themselves also contain metadata that is not marked as metadata but may be identified with machine learning methods (e.g. result pages of search results, table of content pages). Some work has also been done on supervised and unsupervised learning approaches in this area (Liu, 2008).

2 IMPLEMENTATION

The EERQI crawler is based on Nutch (Nutch, 2009), which is an open source web crawler, that is highly configurable and extensible via plugins. It is scalable across CPU clusters by incorporating the Apache Hadoop (Hadoop, 2009) framework.

The following sections discuss the implementation of the search engine for the significant goals mentioned in the introduction.

2.1 Combining Techniques for Best-first Focused Crawling

2.1.1 Crawl Cycle

The Nutch crawler used within this investigation is substantially optimized. The Nutch software itself is not implemented for focused crawling but is extendable in this respect. The crawl is initialized with a seed list: a set of start URLs. Most of these start URLs have been selected from lists of electronic journals in "educational research". These URLs are injected into the Nutch crawl database ("crawldb"), which includes some information about each URL, such as the current status (e.g. fetched or unfetched) and time of last fetch.

Each crawl cycle generates a list of top scoring unfetched URLs or URLs which need to be refetched. These URLs are then retrieved from the WWW and the resulting files are parsed. The URLs and corresponding anchor texts are also extracted and inserted into the link database ("linkdb"). This contains a list of inlink URLs and anchor texts for each URL. The parsed text is indexed if the document meets the Educational Research Document Detection (ERDD) criteria. A partial index is created for each crawl cycle. Duplicate documents are deleted from the indexes ("dedup"). At last, the indexes from each crawl cycle are merged into the final index. The modified status information for each URL is rewritten to the "crawldb". The score for each URL is adapted for EERQI focused crawling ("rescore"). Nutch uses the OPIC (On-line Page Importance Computation) (Abiteboul et al., 2003) algorithm to assign scores to each URL.

2.1.2 Focused Crawling Based on Link Analysis

A basic premise in OPIC and PageRank is (Abiteboul et al., 2003): a page is important, if important pages are pointing to it and important pages should be fetched first and more often. Within the EERQI crawler, we know which pages are important, aka relevant, as soon as we have fetched and analyzed them. These are the pages that have been indexed after being detected as Educational Research Documents (ERD). We must learn to predict, which pages will be important before they are fetched, and follow the most promising paths.

Some samples from the WWW have shown that the ERDs, most often do not link to other important ERD, if they link to anything at all. However, the pages linking to ERDs can be regarded as important pages, because they often consist of tables of content pages for an entire journal volume or year. They will not be indexed but are important in finding links to other relevant pages. It makes sense to use backpropagation for boosting the relevance score of pages which link to ERDs. These pages are comparable to the hubs in Kleinberg's HITS (Hyperlink-Induced Topic Search) algorithm (Kleinberg, 1999). The HITS algorithm assumes that a good hub is a document, that links to many good authorities (authorities are important pages, comparable to ERD). Simply using the above mentioned link importance algorithms (such as OPIC, HITS or PageRank) is not feasible because we will not crawl a significant portion of the WWW and these algorithms do not take into account whether a document is an ERD.

The web may be displayed as a directed graph. Intuitively, an ideal crawl path would retrieve a very high number of ERD and a small number of non-ERD pages. The ratio of ERD pages to the total number of fetched pages should be as high as possible. When considering specific URLs, pages are important, if they link to a high number of pages classified as ERD. Indirect outlinks (outlinks of outlinks) will be considered up to a certain distance. Effectively, the high score of an ERD will be backpropagated to pages linking to it. The resulting score must then be passed on to the outlinks of these pages, until they reach a significant amount of unfetched pages.

We calculate the score based on the ratio of ERD classified outlinks to all outlinks. Ultimately, the total number of ERD classified outlinks was included into the equation but experimental results showed that this did not improve results significantly. Because the total score of each link level should be smaller with growing distance from start level, it must be divided by a constant, here named g, which should fulfill $g > 1$ (experimentally a value of $g = 2$ has proven to yield best results), exponentiated with the link level k. So the score will become weaker, the farther it is propagated. Experiments were used to refine the equation based on the results.

Equation 1 is an heuristic approach newly developed within this project to sum up the score calculations based on backpropagation. It is applicable to a vertical search engine in other fields of interest as well. It has proven to yield optimum results as will be shown later within Figure 2.

$$h_i = \sum_{k=0}^{l} \frac{\frac{c_k}{d_k+1}}{g^k} \quad (1)$$

h_i is score for page i, l is number of link levels, c_k is the number of links of i in link level k that have been classified as ERD, d_k is the total number of links of i in link level k.

2.1.3 Anchor Text Analysis

Up to now, 60,000 anchor texts were analyzed. It may be assumed that words such as "pdf", "full", "article" and "paper" are good indicators of research documents but they do not contain any information about whether the referenced document is about "educational research". The word "abstract" is a good hint, that the referenced document contains only an abstract, which is currently not considered as ERD by the search engine.

SVMlight (SVMLight, 2009) was used to train the anchor texts. SVMlight is a Support Vector Machine based classifier. Single-word anchor texts that are a good indicator of a direct link to research texts ("pdf") obtained almost the same result as single words that would most likely not point to research documents ("sitemap" and "abstract"). It is assumed that this is due to the large number of non-ERD documents (for example research from other fields) that were also linked with potentially promising anchor text words. However, the classifier works well on anchor texts containing typical "educational research" terms, for example "Teacher" received a score of 4.28, "Learning" a score of 4.84.

When training the classifier, not only the anchor texts with direct links to ERD were used, but also anchor texts of indirect links up to a level of three.

An SVMlight score above 0 may be interpreted as a positive hit. The higher the score, the higher the probability of being in the trained class. The maximum score obtained in a list of 30000 samples was 4.89 while the minimum was -4.99. While using this score may optimize the focused crawler, it may also bias the search engine towards documents with "typical" mainstream titles.

2.2 Educational Research Document Detection

Before analyzing how an ERD may be detected, we must first define the term ERD more precisely: An ERD is a digital scientific research document which may be classified within the topic "educational research". It may be for example a journal article, a conference paper, a thesis or a book. An ERD may consist of one or more ERDs as in conference proceedings or entire journals. Abstracts are a part of an ERD but are not considered as a fully qualified ERD.

Educational Research Document Detection may be regarded as a combination of identifying scientific research documents and topical ("educational research") classification.

A large number of publications have analyzed the use of Vector Space Model based algorithms for document classification. Sebastiani (Sebastiani, 2002) provided an overview. These methods may be used for matching new documents with existing categories, such as specific topics (e.g. physics, biology), spam / no-spam etc. The document is represented as a vector. Each dimension of the vector represents a term, the value is a representation of the frequency that the term exists in the document (e.g. tf-idf may be used). When classifying a document, the term vector of the document is matched with the term vectors of the classes. ERDD may be regarded as a binary classification problem, because there is only one class (ERD), or a ranking problem where the documents are sorted by their ERD ranking score.

For supervised learning text classification, a collection of documents is required, which may be used as a training base. This collection should cover all areas of "educational research". A negative collection should be provided as well, which covers documents that should not be considered as ERD, such as research documents from other fields and non-research documents.

The detection mechanism is implemented using the following:

1. A rule based content analysis is used in order to ensure a high probability that the document is a research document. The document must have a minimum text length, it must contain a set of keywords (such as references, abstract) and it must contain references which may be existing in various formats.
2. A number of significant "educational research" keywords must exist in the document. Further work needs to be done to replace or augment this with a vector space model based classifier.

2.3 Metadata

A common Dublin Core based XML format was defined for metadata. The local content base consists of a number of full-text documents and metadata, that have been supplied by various publishers. The full-text content and available metadata was indexed using Lucene (Lucene, 2009).

The Nutch (Nutch, 2009) crawler also uses a Lucene (Lucene, 2009) index. The EERQI indexing plugin was modified to write the full-text content, file format (Internet media type), number of pages (in case of PDF) and the language (as detected by the Nutch LanguageIdentifier) to the index. A list of information about journals was generated by combining the lists of educational research journals from various journals lists with information supplied by EERQI partners. Metadata information in the index was expanded with journal information such as peer-review status, publisher and ISSN. This may later be enhanced with information about the title, authors and date of publication.

2.4 Multilingualism

Multilingualism is of special importance for European search engines due to Europe's diversity of languages. The documents and metadata indexed by our search engine are in fact supplied in several languages. Futhermore the documents themselves may contain more than one language: often the abstract is in English and the rest of the document is in the native language of the authors. In order to provide full multilingualism, it is necessary to use language independent algorithms wherever possible and supply translations for the supported languages.

Our focused crawling scheme based on link analysis is language independent. The ERDD must supply mechanisms for all supported languages.

When submitting a query from the web user interface, the search engine may return results matching the query and related terms in all supported languages. This will be implemented in the very near future of the project using a multilingual thesaurus. To the best of our knowledge this will soon be the first and only search engine implementing this kind of extended multilingualism.

3 RESULTS

At this stage, a prototype search engine has been designed and tested. The primary targets as described in this paper have been implemented.

In order to test the search engine, a number of 100 URLs were randomly selected from the seed list to test the crawler. A mutually exclusive list of 100 further URLs were used to train the anchor text classifier. When crawling, 1000 URLs were generated for each crawl cycle. The crawling alternates between selecting the 1000 best-scoring URLs and selecting the 10 top-scoring URLs for each of the sites from the seed list. This was done to prevent an excessive downgrading of individual sites.

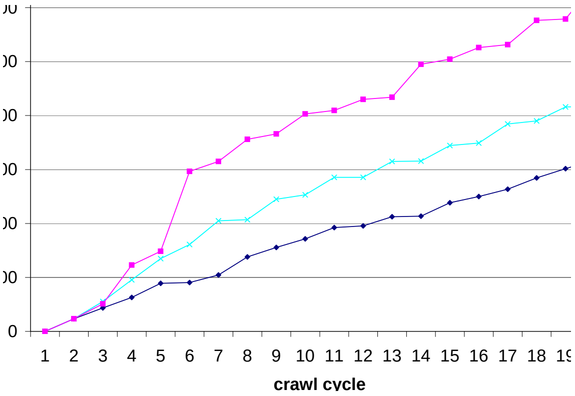

Figure 1: Crawl results with and without optimizations.

The total number of ERDs found for each crawl cycle is shown in Figure 1. Crawling was done with a total depth of 100 crawl cycles. The three lines show the execution of the runs:

1. without optimization,
2. with anchor text optimization: A preliminary train crawl of depth 50 was done with the aforementioned train URLs. When optimizing during the test crawl, anchortexts were rated by SVMlight based on the previous training set,
3. with link analysis optimization by equation 1 : Pages linking to ERD were boosted in score up to a level of three outlink levels and one inlink level. Equation 1 so has proven to be of significant value for intelligent focused crawling and might therefore be used for any vertical search engine crawler.

Using various sources (input from EERQI partners, ERIC database (ERIC, 2009)) a list of URLs was compiled for measuring ERDD. This list included "educational research" documents, research documents from other areas of research, non-research and other documents (such as bibliographies, book reviews etc.). Applying the ERDD mechanism to these documents produced the following results (Table 1):

Table 1: Precision, recall and accuracy ratios in ERDD results.

Precision	Recall	Accuracy
0.73	0.89	0.86

For the four project languages (English, French, German and Swedish) a sufficient number of documents has been gathered from the WWW which is further to be used for testing quality indicators detection within the EERQI project. Results of this detection may be used to enrich document metadata within the search engine.

4 Summary and Conclusion

It has been shown that an advanced vertical search engine collecting full-text and metadata from inhomogeneously structured information sources can be successfully implemented by integrating an intelligent focused crawler. Content classification (ERDD) and metadata extraction have been shown as valuable methods for enhancing search results. Link analysis optimization achieved considerably better results than anchor text optimization or no optimization. Using link analysis, the number of necessary crawl cycles are reduced by at least 50 %, leading to faster results and less use of resources. The EERQI search engine, accessible on the EERQI project website (EERQI, 2009), provides extensive search capabilities within metadata and full-texts. It is the goal of the search engine to gather information about a large number of relevant "educational research" documents and provide access to information about these documents. The first steps have been taken to achieve this goal. A new formula (equation 1) has been developed for focussed crawling.

5 FUTURE WORK

Based on the current implementation, the next stage of the EERQI search engine development will concentrate on optimized content classification (ERDD) and metadata extraction. Further effort needs to be put into metadata extraction from anchor texts and full text. Preliminary tests revealed that a significant number of anchor texts include title, author and / or journal names. This may be combined with metadata extraction from full-texts. The search engine user interface will be enhanced to facilitate ergonomic usability for a number of features, such as clustering and sorting of results as well as complex search queries.

ACKNOWLEDGEMENTS

We kindly thank the partners within the EERQI project and the colleagues at RRZN for their valuable input and support. EERQI is funded by the European Commission under the 7th Framework Programme, grant 217549.

REFERENCES

Abiteboul, S., Preda, M., and Cobena, G. (2003). Adaptive On-Line Page Importance Computation. In *Proceedings of the 12th international conference on World Wide Web*, pages 280–290. ACM. URL: http://www2003.org/cdrom/papers/refereed/p007/p7-abiteboul.html (HTML).

BEI (2009). British Education Index (BEI). URL: http://www.leeds.ac.uk/bei.

Bergmark, D., Lazoze, C., and Sbityakov, A. (2002). Focused Crawls, Tunneling, and Digital Libraries. In *Proceedings of the 6th European Conference on Digital Libraries*.

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., and Kleinberg, J. (1998). Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. In *Proceedings of the Seventh International World Wide Web Conference*, pages 65–74.

EERQI (2009). EERQI project website. URL: http://www.eerqi.eu.

EERQI-Annex1 (2008). EERQI Annex I - Description of Work. URL: http://www.eerqi.eu/sites/default/files/11-06-2008_EERQI_Annex_I-1.PDF (PDF).

ERIC (2009). Education Resources Information Center (ERIC). URL: http://www.eric.ed.gov.

Google Scholar (2009). Google Scholar. URL: http://scholar.google.com.

Hadoop (2009). Apache Hadoop. URL: http://hadoop.apache.org/.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. (2003). Automatic Document Metadata Extraction using Support Vector Machines. In *Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL 2003)*.

Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, pages 604–632.

Liu, B. (2008). *Web Data Mining*. Springer.

Lucene (2009). Apache Lucene. URL: http://lucene.apache.org/.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Nutch (2009). Apache Nutch. URL: http://lucene.apache.org/nutch/.

OAIster (2009). OAIster. URL: http://oaister.org.

Pant, G., Tsioutsiouliklis, J. J., and Giles, C. L. (2004). Panorama: Extending Digital Libraries with Topical Crawlers. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*.

Scirus (2009). Scirus. URL: http://www.scirus.com.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34:1–47.

SVMLight (2009). SVMlight. URL: http://svmlight.joachims.org/.

Witten, I., Don, K. J., Dewsnip, M., and Tablan, V. (2004). Text mining in a digital library. *International Journal on Digital Libraries*. URL: http://springerlink.metapress.com/content/uuuv5md0gm8clrmw/fulltext.pdf (PDF).

Zheng, X., Zhou, T., Yu, Z., and Chen, D. (2008). URL Rule Based Focused Crawler. In *Proceedings of 2008 IEEE International Conference on e-Business Engineering*. ISBN: 978-0-7695-3395-7, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4690611&isnumber=4690569 (PDF).

Zhuang, Z., Wagle, R., and Giles, C. L. (2005). What's there and what's not? Focused crawling for missing documents in digital libraries. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*.