Capturing Contexts for Web Filtering in the Humanities

Haakon Lund[1], Birger Larsen[1], Rasmus Erik Voel Jensen[1],
Anders Ardö[2], Koraljka Golub[2] and Peter Ingwersen[1]

[1]Department of Information Studies
Royal School Library and Information Science
Birketinget 6, DK-2300 Copenhagen S, Denmark
{hl,blar,rvj,pi}@db.dk

[2]KnowLib Research Group, Department of
Information Technology, Lund University
P.O. Box 118, SE-221 00 Lund, Sweden
{anders.ardo,koraljka.golub}@it.lth.se

1. INTRODUCTION

The web search engines available today need improvement in supporting the user in her selection of relevant scientific information resources. Initiatives like *CiteSeer* and *Google Scholar* are trying to overcome this problem by using more elaborate processes for selecting documents for inclusion in specialized scientific search engines. Other developments have been the establishment of domain specific web portals based on manual selection and annotation of web resources. Most of the implemented solutions have been based on building web portals covering science, technology and medicine, but portals within humanities are few and far between.

The development of domain specific web portals is a time consuming and costly undertaking with respect to human resources. This includes an intellectual evaluation of the usefulness of the individual resources in relation to the scope of the web portal. With the ever-increasing amount of heterogeneous web content a selection process based on human evaluation is not likely to become less resource demanding. A solution could be the development of automatically generated web portals based on algorithmic context analysis for the selection of web resources.

This issue is investigated in the Humanities Portal project. The purpose of the project is to develop more precise methodologies and tools for automated selection and retrieval of scientific information resources within the humanities, foremost in Danish, but also in a multilingual context. This paper presents some of our initial considerations about context-based filtering of web resources, and discusses the implementation of a web crawler aimed at establishing a test collection for empirical experiments with filtering and classification of humanistic web resources. Context is here understood as features of web documents, such as outlinks, their anchors and academic references and their contextual features in the citing document.

2. CONTEXT TYPES

According to Ingwersen & Järvelin (2005, p. 283; Ingwersen, 2005) one may distinguish between six fundamental kinds of context associated with Information Seeking and Retrieval (IS&R), out of which five are nested. The first layer in this model of context consists of intra-object structures, such as words, sentences, paragraphs, etc. Inter-object structures, such as links, form the second layer of context. Further out into the contextual stratification scheme the model depicts interaction, such as IR interaction or other activities connected to person-object relationships, for instance authoring web documents. A fourth contextual type consists of the actors and their cognitive-emotional characteristics, such as knowledge levels, perceived tasks and experiences. This constitutes the socio-systemic and organizational context. The fifth and last nested kind of context is defined as the techno-cultural- and politico-economic infrastructures of society. A sixth kind of context associates to all the former: the temporal dimension.

In the Humanities Portal project we are concerned with web objects of a humanistic nature, i.e., the contexts within and between web pages denoted 'intra-object structures' and 'inter-object structures' in Ingwersen and Järvelin's model. In addition, we seek to capture the contexts and situations of the *authors* of such objects. This implies that we observe the results of the interaction that takes place during generation of web documents. Outlinks and their anchors are typical features that point to this kind of context. Also bibliographic references – and the text passages surrounding the outgoing reference in the citing text pointing to other literature – may be exploited in this sense of context. The passages linked or referred to in the cited documents are to be involved at a later stage. The usability testing of web objects for information retrieval and acquisition purposes is not considered in this approach.

3. PUBLICATION PATTERNS

A small convenience sample was collected to investigate the online publication patterns of researchers in selected areas within science and humanities (Voel Jensen, 2005). The scope of this pilot was to identify to what extent researchers publish their research on the institutional web pages and to identify the validity of institutional web pages as a resource for harvesting scientific publications for web portals.

Web sites from science and humanities at the University of Copenhagen and the University of Roskilde were analyzed and types of publication surrogates from the first 10 scientists mentioned in the institutional directory were collected. Data was then classified into 4 categories according to type of document representation available:

Full text = articles in full text available
Abstact = abstract available
Bibliograhy = list of references available
Negative = nothing available

The existences of Researcher CVs are not included in the data collected.

Table 1. Web pages from 60 scientists were analyzed in total. The table shows the number of researchers' web pages found for each type of information and is not an indication of the actual number of publications referred to. (UoC = University of Copenhagen, RUC = University of Roskilde)

Department	Negative	Biblio-graphy	Abstract	Full text
Linguistics, UoC	8	2	0	0
History, UoC	4	6	2	0
History, RUC	3	7	0	0
Biology, UoC	7	3	2	0
Physics, UoC	5	5	2	2
Physics(+math), RUC	0	10	2	2
Total	27	33	8	4

In a few cases it was possible to identify more than one type of document representation or document overlap. Web pages from 4 researchers provided access to full text, abstracts and bibliographies, and 4 researchers provided access to abstract and bibliography.

Table 2. Publication patterns by faculty.

Faculty	Negative	Biblio-graphy	Abstract	Full text
Humanities	50%	43%	7%	0%
Science	40%	40%	7%	13%

Data were collected from a relatively small sample of only 60 researchers and as seen in tables 1 and 2, no full text publications were available from Humanities. The available full text articles from Science were all from Physics. This may be due to extensive document repositories within Physics a tradition not common in biology and humanities. Most of the researcher's web sites at RUC (85%) are *linked* to a central database of publications resulting in availability of bibliographies. This is not the case at the UoC (65%) even though a database does exist. This result could indicate that a harvesting of data should include the university publication databases, and not only rely on data available from the researcher's web sites.

4. EXTRACTION OF WEB PAGES

To automate the collection of scholarly web resources, patterns for identifying relevant resources has to be established. The underlying problem is how to separate scholarly information form other non-relevant types of information. Obviously identification of the originating URL of a web page can be used to identify web resources published from known research institutions. However there is no evaluation of the scientific relevance of the captured web pages and there is a risk of not capturing web resources originating form other relevant resources. Nicholson (2003) presents 4 different models for identifying scholarly web pages. According to Nicholson the *Classification Tree Model* is the most successful achieving an accuracy of 96.00% in identifying academic web pages from non-academic. Nicholson's *Classification Tree Model* uses 13 criteria for identifying the type of web page:

1. Number of references in the text
2. Average word length
3. Existence of references to "Table 1" or "Figure 1"
4. Number of times a traditional heading appeared on the page (such as Abstract, Findings, Discussion, etc.)
5. Number of times phrases such as "published in," "reprinted in," etc. appear
6. Academic URL
7. Ratio of total size of images on page to total size of page
8. Number of misspelled words according to Dr. HTML[1]
9. Number of words in the meta keyword and dc.subject meta tags
10. Average number of punctuation marks per sentence
11. Average sentence length
12. Number of sentences in the document
13. Commercial URL.

(Nicholson 2003 p. 1087)

The web resources identified by Nicholson only cover web pages in HTML leaving out other types of published documents such as PDF files etc. According to a study by Jepsen et al. (2003) the use of PDF files in web publication correlates with scientific content. Therefore an extension of the *Classification Tree Model* to include PDF-files is essential. From a multilingual perspective the model is in some cases only looking at attributes relevant for resources in English as criteria 4, 5 and 8 (in Dr. HTML spell checking resources in Danish is not available). The use of criteria 13 is not a viable measure for identifying the type of URL for non-US domains. However the *Classification Tree Model* does provide a frame work for developing a multilingual model to identify scholarly web resources. To select the most relevant web pages, i.e., in the proposed project scientific web resources from Humanities, we plan to extend the model with domain specific criteria.

5. A WEB CRAWLER MODEL

In addition to criteria for filtering out academic web resources a focussed web crawler is needed. The key to a successful focused Web crawler is to enable the crawler to select the most relevant links to follow in order to find the most relevant pages (with respect to the focus topic). A model of the crawler architecture we plan to use is shown in fig. 1 below: A set a *start pages* is passed through a *URL policy filter* and a *scheduler* controls the *harvesting* and a *database* of harvested documents. Inspired by work done at Lund University (Ardö 2005) we plan to evaluate the relevance of a page for a specific topic subject area by automated subject classification. In this approach the core algorithm for automated subject classification in the *content policy filter* (fig.1) is based on matching of terms from a *topic definition* with the text of the document to be classified. Each time a match is found, the document is assigned the corresponding class, and awarded a relevance score, based on which term is matched (single word, phrase, Boolean)

[1] http://www2.imagiware.com/RxHTML/

(*weight[term]*), and the part of the Web page in which the match is found (*weight[loc]*). A match of a phrase (a number of words in exact order) or a Boolean expression (all terms must be present but in any order) is made more discriminating than a match of a single word. A list of suggested classes and corresponding relevance scores (S) is produced using the following algorithm:

$$S = \sum_{locs} \left(\sum_{terms} (freq[loc_j][term_i] * weight[term_i] * weight[loc_j]) \right)$$

The single scores are summed to make the final score (S) of the document. If that score is above a cut-off value, the document is saved in the *database* together with a (list of) subject classification(s). In the future this might be replaced with a more dynamic method that makes it easier to adopt to new term lists or changes in old ones. (Ardö 2005). In the Humanist Portal project we intend to investigate how to exploit the contextual features mentioned in section 2 to refine the topic definition and to use them to filter out humanistic documents from other documents. We also plan to study how to exploit the contextual features to improve *clustering*, *ranking* and *cleaning* of the final database.

6. CONCLUSION

In this article, we report on the preliminary design considerations in relation to creating a web portal for the Humanities. The task of creating web portals can be a daunting task and combined with the increasing amount of heterogeneous resources published on the web tools for automatic creation of web portals is inevitability. The proposed project is working in a multilingual environment were a substantial number of publications are believed to be publish in a number of languages. To identify the publication patterns within the Humanities regarding web publication data were collected from the official web pages of researchers from two Danish universities. The data showed that none of the researchers from Humanities had any full text documents available, but some do publish bibliographies. In a number of cases as links to central publication data bases, signifying that resources like these must be included in a web portal. To verify these preliminary finding the analysis has to be applied on a more extensive data set.

A issue in harvesting scholarly, domain specific web resources is the identification of relevant web pages. To identify scholarly web pages from other non-scientific web pages is the first step and the proposed *Classification Tree Model* would have to be revised in order to identify multi lingual web pages and documents publish in other formats than html.

7. ACKNOWLEDGMENTS

We wish to thank the Swedish Agency for Innovation Systems and the Danish Ministry of Culture (grant no. A2004-06-028) for partly funding this study.

8. REFERENCES

Ardö, A. (2005): Focused crawling in the ALVIS semantic search engine. In: *2nd European Semantic Web Conference, 29 May – 1 June 2005, Heraklion, Greece*, p. 19-20.

Ingwersen, P. (2005): Selected variables for IR interaction in context: Introduction to IRiX SIGIR 2005 Workshop. *IRiX ACM-SIGIR 2005 Workshop Proceedings,* p. 6-9. Available at: (http://irix.umiacs.umd.edu/)

Ingwersen, P. & Järvelin, K. (2005): *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer. 463 p. (In press)

Nicholson, S. (2003): Bibliomining for Automated Collection Development in a Digital Library Setting : Using Data Mining to Discover Web Based Scholarly Research Works. *Journal of the American Society for Information Science and Technology*, 54(12), p. 1081-1090.

Thorlund Jepsen E., Seiden, P, Ingewersen P., Björneborn L. and Borlund P. (2004): Characteristics of Scientific Web Publications: Preliminary Data Gathering and Analysis. *Journal of the American Society for Information Science and Technology,* 55(14), p. 1239-1249.

Voel Jensen, R.E. (2005) Undersøgelse af onlinepubliceringsvaner [Investigation of online publication patterns]. Danmarks Biblioteksskole. (Internal report, in Danish)

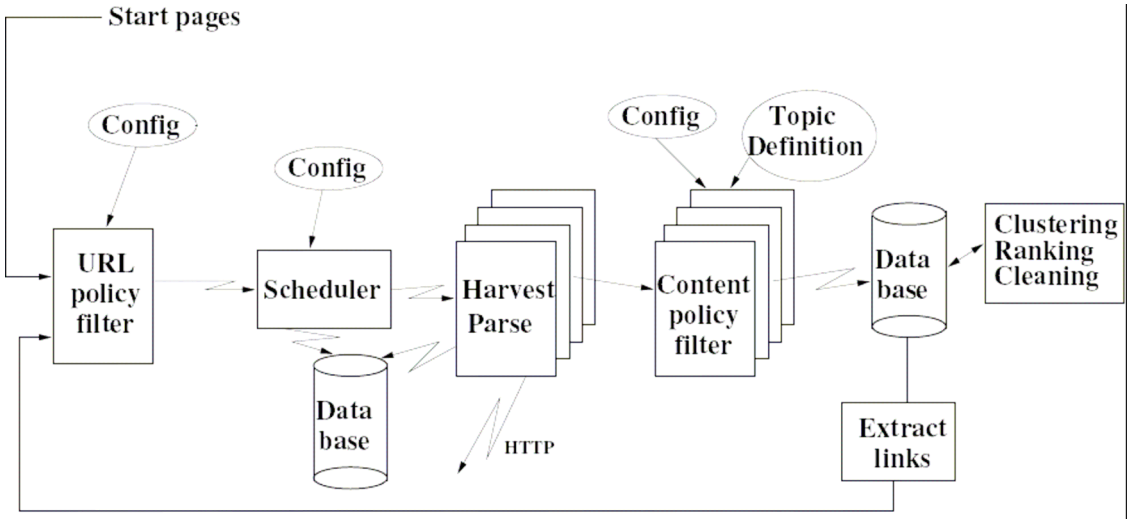

Figure 1. Focussed crawler architecture