Automatic Web Resource Discovery for Subject Gateways

Konstantinos Zygogiannis[1], Christos Papatheodorou[2], Konstantinos Chandrinos[3], Konstantinos Makropoulos[4]

[1]Dept. of Communication Systems, Univ. of Lancaster, Lancaster, UK
[2]Dept. of Archive & Library Sciences, Ionian University, Corfu, Hellas
[3]Institute of Informatics & Telecom., NCSR Demokritos, Athens, Hellas
[4]Division of Applied Technologies, NCSR Demokritos, Athens, Hellas
[1]k.zygogiannis@lancaster.ac.uk, [2]papatheodor@ionio.gr,
[3]kostel@iit.demokritos.gr, [4]cmakr@nh.gr

Abstract

Subject gateways have been heralded as the qualitative answer to information overload and the meagre ability of current generation search engines to provide context-sensitive query answering. However, human maintenance proves problematic as soon as the domain broadens slightly. In this paper we report on an attempt to leverage the maintenance problem of a successful subject gateway for chemistry, by providing automatic tools. The tools collect and rank quality Web resources, utilizing the existing ontology of the gateway. Evaluation of such a Web search mechanism against classic web queries to popular and topic-specific search engines gives promising results.

1 Statement of problem

Recent years have seen a number of efforts attempting to leverage information overload. Although the most obvious aspect of information overload is quantity, the problem of information quality has started to receive special attention, particularly when treating the Internet as a source of educational or professional information. Information quality attempts to answer the question: how can one select from "everything" only those information items that meet one's information needs and at the same time carry a certain validity or authority?

The immediate reaction to this was the design and implementation of hierarchical directories that attempted to capture the real-world subject hierarchy and assist the user in narrowing down the number of resources he/she had to consult. Two critical virtues of such efforts, soundness (the fact that all links pointing to the intended resources are valid) and completeness (nothing equally or more interesting exists outside these links) were vanishing on a daily basis. Soundness, although not guaranteed at 100%, is often supported by automated link-checkers, while completeness or near completeness has become an issue of commercial competition. The larger the army of human indexers, the higher the completeness of the directory is.

To address content quality, long-standing traditions and techniques coupled with novel technology have given rise to a new type of information gateways, the so-called "quality-controlled subject gateways". These are "Internet services which support systematic resource discovery" (Koch, 2000). Important aspects of a quality-controlled subject gateway are that it provides access to

resources that are mostly on the Internet along with descriptions of these resources and the ability to browse through the resources via a subject structure. In the following paragraphs the term "subject gateway" is to be interpreted as a quality-controlled Internet resource catalogue, with descriptive documentation and a hierarchical browsing facility.

Subject gateways employ one or more quality criteria for selecting resources. The employed criteria manifest themselves in the resource selection, in the hierarchy decisions and/or the descriptions that accompany every link to a resource. Subject gateways are hard to maintain, primarily because the criteria for content inclusion are not necessarily compatible with the indexing mechanisms provided by Web search engines. Although it is relatively simple for the administrator of a subject gateway to construct automated Web queries and augment the gateway with the links returned, it is often the case that the results contain marginally relevant or even irrelevant answers. This is not alleviated either by the so called "relevance ranking" performed lately by the search engines, although a number of algorithms have been proposed (Yuwono & Lee 1996; Carrière & Kazman; Kleinberg 1998; Page et al., 1998). Relevance ranking within the result set is often used because there is no hierarchical structure of concepts to signify the context when issuing a query. Relevance metrics are tuned to popularity and cross-reference, rather than quality, which is of course a subjective judgement and hence cannot be automatically computed by a search engine.

Although the difficulty of maintaining a subject gateway, there are continuing efforts to provide automatic "intelligent" software components that will attempt to employ "quality" criteria while they are crawling the Web for resource discovery. Personalization techniques learn, model and utilize the subject gateway users behavior to discover interesting information resources and customize personal portals (Anderson & Horvitz, 2002). The Focused Crawler system (Chakrabarti et al., 1999) bounds its crawl to find links that are relevant to web pages indicated by a user and tries to identify the pages that are great access points to many relevant pages. Ontologies address the semantics problem of the query words issued to the search engines. An ontology (Noy & McGuiness, 2001) is a formal and explicit description of concepts in a domain of discourse as well as their properties and features and the restrictions on these properties. Concepts are usually organized hierarchically. Ontologies include machine-interpretable definitions of basic concepts in the domain and relations among them and define a common vocabulary for sharing information in a domain. Many intelligent search engines use ontologies to organize web pages according to their topical hierarchical structure (Tiun et al., 2001; Tandudjaja & Mui 2002).

In this paper we report on work contributing to the automation of maintenance of a quality-controlled subject gateway. Our work aims to provide a subject gateway administrator with a system, which aids him in discovering new resources using the controlled vocabulary offered by a domain specific taxonomy. In doing so, we expect that the quality criteria employed by humans in crafting taxonomical hierarchies will be maintained in the newly discovered resources. Our experiments were based on a well-established and popular subject gateway on Chemistry, but the design and implementation of the mechanisms for automatic resource discovery is domain independent.

2 The Chemistry Information Retrieval

"Information Retrieval in Chemistry" subject gateway (http://macedonia.chem.demokritos.gr) has developed a 3 level taxonomy followed by a controlled vocabulary for representing the chemical knowledge and related domains. For each level there is a set of codes.

- 1st level: Chemistry or Chemistry Related scientific domains (two-digit codes for example "Chemistry" code=01 or "Energy" code=06)
- 2nd level: Chemistry or Chemistry Related sub-domains (Biochemistry, Organic Chemistry, Food Chemistry, etc., three-digit codes, for example "Biochemistry" code=050 or for the Chemistry related "Biotechnology" code=200)
- 3rd level: Information Resources (Books, Journals, Databases, Conferences, Mailing lists, Academic Servers, News/newspapers/magazines, etc., two-digit codes for example "Database" code=15)

In view of the above, the combination of codes defines a query to the subject gateway (i.e. 01-050-15 means "Biochemistry: databases").

The administrator maintains the subject gateway by performing a search into popular search engines, extracting the results and finally selecting the best. Given the current size of the gateway a full cycle through the taxonomy can be achieved twice a year on a part-time basis or up to four times a year on a full time basis. We designed and implemented a set of automatic mechanisms that allow the administrator to go through this cycle much more rapidly, by utilizing the implicit ontology to construct queries to popular search engines. We re-rank their results minimising the number of links requiring visual inspection for quality assurance.

3 The proposed architecture

We designed and implemented a client/server architecture using Java for content collection and processing and HTML for the Web interface. The proposed system consists of three modules. These are the searching module, the ranking module and finally the end-user interface.

3.1 Searching

The administrator issues a query through a web form, using the controlled vocabulary terms provided by the taxonomy. The query is re-formatted appropriately and propagated to a list of predefined popular search engines including a large number of Chemistry related search engines and portals. Moreover he/she can choose the number of the results that will be analysed from each response set, including "all results". In this version, we have used hard-coded wrappers to extract the links pointing to content from the returned results. The searching procedure removes duplicates and invalid URLs and stores the requested number of links in a database. Since the search engines rank their results, the stored links are the top ranked from each response set.

3.2 Ranking

The Ranking servlet extracts the search results from the database where they were stored from the Search procedure, processes them and finally stores the ranked results to a different database. During its operation the algorithm methodically examines all the various links provided and uses regular expressions, substring matching and the edit distance for errors and misspellings since the domain contains a lot of technical terminology.

In particular, if the exact taxonomy term has been found in a specific link, then the algorithm marks this link with a score accordingly to (i) the frequency of the taxonomy term as well as (ii) its position in the retrieved page (e.g. if the taxonomy term is found in the meta tags instead of the main text, for example in the title of the retrieved page, then the page takes a bonus). Moreover if a taxonomy term consists of more than one word, then the algorithm combines the above-

mentioned criteria with the proximity of the taxonomy term words in the retrieved page. If the proximity of two words or phrases is greater than a maximum allowed distance in the retrieved page, then the algorithm considers that the taxonomy term does not exist in the retrieved page. For instance, for a given example of a taxonomy term consisted of two words, it could be stated that these words should occur within four words. In the case of a relevant matching, the algorithm ranks the retrieved page with respect to the (i) edit distance that the taxonomy term has with the words in the retrieved page (e.g. if taxonomy term is 'journal' and a word in the retrieve page is 'journalism' then edit distance equals to 3) and (ii) above-mentioned three criteria (position, frequency and proximity) of the similar words in the retrieved page.

3.3 User Search algorithm and user interface

The ranked results can be browsed by the users through a web form. The users can select at a query one or more Chemistry related sub-domains (Educational, Databases, Glossaries, Journals), but only one domain (Organic, Inorganic, Physical, Analytical, and Environmental Chemistry). Furthermore they can select a sub-set of the search engines used by the searching module of our system to collect links. The query is performed offline, on the previously aggregated, ranked and potentially approved links. The results are returned in the form of clickable HTML links and their ranking is mentioned.

Given the positive results of the work so far, we intend to experiment by allowing the user, after appropriate warning, to send queries to the selected search engines at real time, when the existing results do not satisfy the original query. This will allow us to define a trade-off between the delays of answering with the topicality of the real time results.

4 Evaluation

An early evaluation test was performed with the first prototype. Queries were posed to four search engines, including two Chemistry related, Yahoo and Google, as well as to the proposed automation mechanisms. During this evaluation, each search engine was queried ten times for ten different subjects. The queries issued were: Organic chemistry journal, Chemical suppliers, Chemical magazines, Chemical portals, Chemical DNA structure, Spectroscopy, Nuclear Magnetic Resonance, Apolipoprotein Antigen, Acetyloleolinesterace, Ion chromatic analysis system. The subject gateway administrator browsed the results and selected the relevant answers summarised in the following table, after removal of duplicates.

Table 1: Relevant Answers for a Sample of 100 Queries

Query	1	2	3	4	5	6	7	8	9	10
ChemIndustry.	18	19	23	20	32	42	16	32	29	09
Chemie.de	34	15	12	29	28	30	19	41	16	13
Google	23	28	22	11	13	19	6	9	0	1
Yahoo	19	24	14	12	10	6	2	17	3	0
Total	56	53	43	42	41	23	29	51	14	16
Our system	69	73	78	69	71	49	53	74	32	43

The *Total* row of Table 1 accumulates the overall number of correct results. This was produced by adding the number of correct results returned by each search engine excluding duplicate findings than may occur during this addition. Table 1 shows that the designed gateway maintenance

method produces in total, more relevant results than the rest of the engines do together. Also, our system performs best for highly specific queries (e.g. Acetyloleolinesterace or Ion chromatic analysis system). It that cases it returns almost three times more correct results than all the engines. This is because our system utilizes all available results, not only the first one hundred.

5 Conclusions and future work

In this work we have tried to provide automated tools for the assistance of a subject gateway administrator in maintaining the freshness of its content. We did so, by exploiting the implied ontology from the taxonomic representation of the domain. However, our tools are at no point domain specific, so they could be easily re-used for different domains and gateways. The evaluation results are a strong indication that our system could easily become the automated assistant of the subject gateway administrator, collecting appropriate material in minimal time and reducing it to a tractable quantity that would allow him to keep the gateway up-to-date.

A number of implementation choices will be revised in a second version. For example, the use of open queries with term matching against the ontology via thesauri or the hard-coded vs. automatic wrapper induction for the various search engines. Additionally, we intend to explore learning techniques from the content of links already approved and included in the directory structure so as to achieve a more thorough modelling of each concept and utilize query expansion for ranking.

References

Anderson, C.R. & Horvitz E. 2002. Web Montage: A dynamic personalized start page. In *Proc. 11th Intl. World Wide Web Conference*. ACM Press.

Carrière, J. & Kazman, R. WebQuery: Searching and visualizing the web through connectivity. Retrieved February 7 2003, from http://www.cgl.uwaterloo.ca/Projects/Vanish/webquery-1.html

Chakrabarti, S., van der Berg, M. & Dom, B. 1999. Focused crawling: A new approach to topic-specific web resource discovery. In *Proc. 8th Intl. World Wide Web Conference*, 545-562.

Kleinberg, J. 1998. Authoritative sources in a hyperlinked environment. In *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*.

Koch, T. (2000). Quality-controlled subject gateways: Definitions, typologies, empirical overview. *Online Information Review*, 24(1), 24-34.

Noy N.F. & McGuiness, D.L. 2001. Ontology Development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. Stanford Digital Libraries Working Paper.

Tandudjaja, F. & Mui, L. 2002. Persona: A contextualized and personalized web search. In *Proc. 35th Hawaii Intl. Conference on System Sciences*. IEEE Press.

Tiun, S., Abdullah, R. & Kong, T.E. 2001. Automatic topic identification using ontology hierarchy. In *Proc. 2nd Intl. Conference on Intelligent Text Processing and Computational Linguistics*, LNCS 2004, (pp. 444-453). Springer-Verlag.

Yuwono, B. & Lee, D.L. 1996. Search and Ranking Algorithms for Locating Resources on the World Wide Web. In *Proc. 12th International Conference on Data Engineering*, 164-171.