Focused Crawling: Experiences in a Real World Project

Antonio Badia, Tulay Muezzinoglu and Olfa Nasraoui
CECS department
University of Louisville
Louisville, KY 40292 USA
[abadia|tulay|olfa.nasraoui]@louisville.edu

Categories and Subject Descriptors: H.4.m Information Systems:Miscellaneous

General Terms: Algorithms

Keywords: crawling, topic, information retrieval, thesaurus.

1. INTRODUCTION

Focused crawling is the act of examining a collection of hyperlinked documents (i.e. the Web) to find out those that are about a certain topic ([2, 1]). In contrast, general (unrestricted) crawling examines the whole collection, gathering some information (keywords) about each document.

Here we report on our experiences building a focused crawler as part of a larger project. The National Surface Treatment Center (NSTCenter) is an organization run for the U.S. Navy by Innovative Productivity Inc. (IPI) a non-profit company that provides novel technology-enhanced services and solutions for National Defense, business, and work force customers. The NSTCenter web site was created with the goal to become a premier forum for Navy officers, independent consultants, researchers and companies offering products and/or services involved in the process of servicing Navy ships. In order to help generate content, we developed a focused web crawler that searched the web for information relevant to the NSTCenter. The team has developed a crawling system that achieves significant precision.

Focused crawling has attracted considerable attention recently ([1, 4, 5, 2, 3, 6]). Most methods use primarily link structure to identify pages about a topic, or combine several measures from text analysis and link analysis to better characterize the page. [5] proposes a method similar to the one used here, in that a knowledge structure (an ontology) is used to identify relevant pages. We use a thesaurus, which does not have as much information but is much easier to build and maintain.

Focused crawling on the real web can be extremely difficult for several reasons. First, the concept of topic is not formal or formalized (and perhaps not formalizable). As a consequence, the relationship of being *about* a topic, already difficult to determine, is even more difficult. Second, on a networked collection, the network itself is used to determine page aboutness, on the assumption that a page content can be partially determined by its network topology. However, *topic drift* and other problems make this assumption work only up to a point. The third difficulty in focused crawling is that pages may not be the correct unit of work. There has been some research on the fact that sometimes a page is too coarse a unit, while in some cases, pages may be too *fine*-grained. Finally, another important difficulty is the problem that most web pages are *dirty* from the point of view of content, that is, they either have no real content or, besides their main topic, they also contain other information that is only partially (or not at all) related to that topic.

2. OUR APPROACH

Our solution involves starting with a simple (but efficient) IR approach, by looking at pages that have certain words. To ensure good recall, we use a web search engine with broad coverage in order to cast a wide net (Google) and then filter the obtained results in order to improve precision.

The algorithm proceeds in four phases:

Harvesting phase: in this first phase, we gather pages. As stated, we use Google in order to increase recall, even if at the cost of precision (later on, we will work on precision alone by filtering the pages). The result of several searches in the search engine is used to fill up a queue of pages. First, we need to choose keywords to start the Google search. Our heuristic was to use high-level thesauri words -such words tend to be more general and hence increment recall. However, we found out that some care was needed to combine the words. Too many words tend to lower recall on an exponential scale: 2 or 3 words will bring tens of thousands of pages, 5 or more keywords will only bring hundreds (even less if some words are technical). Our solution uses a large number of searches, each one started with only 2 or 3 words, and then picks only the top n pages from each search. Besides Google, we use two other sources of information, two dynamically maintained lists, one of sites and one of hubs (see later for the maintenance strategy).

Pre-processing phase: in this second phase, we try to discard non-content pages and duplicate pages. We also check if a page in the queue is already in our database. This is due to the fact that we expect this search to be run regularly, and therefore we expect many pages to be retrieved that are already known to the system. It is a serious challenge to determine whether two pages have the same content, in order to avoid redundancies in the result. This issue has also been attacked in previous research, but only to a limited extent. For now, we simply use the URL and date-last-modified to check if we are revisiting exactly the same page and there have been no changes since the last time. A checksum on the text is also used to detect two pages that are verbatim copies of each other; however,

WWW 2006, May 22–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

highly related pages pass this test. Also at this point, we get rid of forums and blogs. The decision to do so was taken given our need for authoritative sources. Detecting blogs and forums is done through an extremely simple test: we simply check if the string “forum” or “blog” are present in the page’s URL. While this is trivial, it happens to work surprisingly well. Detecting hubs, on the other hand, turns out to be a daunting task due to the issues with page content and presentation mentioned above. Currently, we count the number of links to an external site in the page and divide by the number of words in the page after taking away all HTML (including the anchors themselves) and any stopwords (in an IR sense).

Filtering phase: in the third phase, we decide whether a given page is about our topic. We start by cleaning up the page (creating a text-only version). The core of our algorithm is the comparison of the page’s text with the thesaurus. We carefully edited a thesaurus and match words in it against words in a page. We structured the domain after interviews with domain experts and review of relevant material. Once a basic structure was agreed upon, it was “coerced” into the thesaurus. The coercion was needed because most thesauri support only some basic functionality (relations), while the domain (like most domains) required more fine-grained divisions. For instance, since our general theme (corrosion on ships) was quite wide, we divided it into aspects or *facets*, a basic idea borrowed from Information Science. Thus, we divided the topic into areas like *ships* (the subject), *methods* (used in combating corrosion), *materials* (used by those methods), *people* (involved in some aspect on the task: chemical engineers, consultants, etc.), *organizations* (makers of products, providers of personnel, or otherwise involved in the effort). As a result, the thesaurus was structured as a *forest* or list of trees. Each tree corresponds to a facet and contains a taxonomy inside; since each tree has *topic coherence* (that is, all the words under the root are closely related), we identified a topic (or subtopic) with one such tree in the thesaurus. For the matching process, we noticed that, since our topic was quite wide, most pages would only match a small percentage of the words in the thesaurus. Therefore, we weighted the matches according to some simple heuristics. First, each page was matched against each thesaurus subtree representing a facet (topic) separately, and a score obtained for each such match. Second, frequency of words was not counted heavily, but diversity of words was. Third, words in lower levels of the thesaurus were given higher weight than words at the higher level, due to the fact that they usually are narrower in scope. Finally, the matching process was slightly modified by the addition of *negative* words and expressions (n-grams), words and expressions that we did want to avoid. Finally, we round up the score of a page by using its URL and links to it. For links to the page, we score an *anchor window* against the thesaurus; for a URL, we build a list of individual words, disregard the site name and score the remaining words. We point out that we expanded the thesaurus with numerous entries, especially proper nouns, obtained from documents, already captured web pages, and other sources.

Post-processing phase: on the final phase, we update our list of hubs and sites by counting, for each hub or site, the number of pages found to be relevant. If the number was above a threshold, we kept the hub or site; otherwise, we disregarded it.

Experimental Evaluation: the standard measures of evaluation for a web crawler are the ideas of precision and recall. However, it is very difficult to assess either one on the Web. Instead, we analyzed our *Google-relative recall* as follows: we ran additional searches on Google to bring more pages to the program. For each search, we run our program in the same way and inspected the final results. We found out that there was a plateau on the number of relevant pages after a certain number of searches (i.e. more searches did not bring more relevant material). With respect to precision, we resorted to the same methods as previous IR research: we used human subjects to judge the quality of our search. On an experiment with 4 volunteers and a random sample of pages, the results were highly encouraging, with a total precision of 83%. In contrast, the precision of Google, as measured by the people in our team, never reached that high (even in the first page of results), and dropped precipitously after the first page.

3. CONCLUSION AND FUTURE RESEARCH

During our experience designing and building a focused crawler, we have found that working on real web pages creates an engineering challenge and a conceptual challenge. On the engineering level, many practical considerations outside the scope of pure research must be tended to. On the conceptual level, determining if a certain page is about a given topic quickly leads to deep questions which are difficult to answer: what exactly is a topic? How do we measure aboutness? In this sense, two important avenues of research have suggested themselves after this experience. First, it is important to determine ways to formalize (or at least approximate) the idea of topic. Second, most approaches consider the text of a page from an IR perspective, i.e. as a bag of words. However, there is clearly more to examining a text than this. There are some challenges that simply cannot be met from this perspective. It is necessary to introduce Information Extraction (IE) and Query Answering (QA) techniques into web crawling in order to achieve real relevance. Our future research pursues these two lines of inquiry.

Acknowledgments: The first and second authors were supported by a grant from IPI. The third author was supported by a NSF CAREER Award IIS-0133948. The first author was supported by a NSF CAREER award IIS-034755.

4. REFERENCES

[1] C. Chung and C. L. A. Clarke *Topic-oriented collaborative crawling*, in Proceedings of the 2002 ACM CIKM, pages 34-42.

[2] S. Chakrabarti, M. van den Berg, and B. E. Dom *Focused Crawling: A New Approach to Topic-specific Web Resource Discovery*, Computer Networks, 31(11-16), 1999, pages 1623-1640.

[3] S. Chakrabarti, M. Joshi, and V. Tawde *Enhanced Topic Distillation using Text, Markup Tags and Hyperlinks*, in Proceedings of the ACM SIGIR, 2001.

[4] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles and M. Gori, *Focused Crawling Using Context Graphs*, in Proceedings of VLDB 2000, pages 527-534.

[5] Ehrig, M. and Maedche, A. *Ontology-Focused Crawling of Web Documents*, in Proceedings of the ACM Symposium on Applied Computing, 2003.

[6] J. Hou and Y. Zhang, *Effectively Finding Relevant Web Pages from Linkage Information*, IEEE TKDE, 15(4), July/August 2003, pp. 940-951.