Language Specific and Topic Focused Web Crawling

Olena Medelyan[1], **Stefan Schulz**[1], **Jan Paetzold**[1], **Michael Poprat**[12], **Kornél Markó**[12]

[1]Freiburg University Hospital, Department of Medical Informatics, Freiburg, Germany
[2]Jena University, Computational Linguistics Research Group, Jena, Germany

1. Introduction

The Web has been successfully explored as training and test corpus for a variety of NLP tasks ([8], [2], [6]). However, corpora derived from the Web are usually inconsistent and highly heterogeneuos in their nature, which is normally counterbalanced by extending their size to billions of words. We assume that web crawling that takes into account domain and language represented by the content of the webpages would allow to acquire huge high-quality corpora. This would bring additional benefits for corpus based approaches to NLP.

Local Web collections are usually created by crawling the WWW starting with few seed URLs. The crawler fetches each webpage, follows its outgoing links and repeats the fetching process recursively[1]. Focused crawling, proposed by Chakrabati ([4]), is designed to narrow the acquisition to Web segments that represent a specific topic. The main assumption is that webpages on a given topic are more likely to link to those on the same topic. Probabilistic classifiers are used to prevent crawling of unrelated websites which would result in a deviation from the specific topic. Such analysis of the Web's link structure became the state-of-the-art in focused crawling ([1], [5], [10]). Only few approaches are known for language specific crawling, and even here link analysis is the underlying strategy ([9]). Our presumption is that focusing on a specific language and domain area is more precise when the content of the documents is taken into account.

2. Content-Based Focused Crawling

Our approach is performed in two steps. Firstly, we create a list with topic and language specific seed URLs. To do this we use a training document collection in the required language and domain area. From these documents we extract ordered non-stopword n-grams with the highest TFxIDF values, which are assumed to be specific to the given domain. We use these n-grams as focused queries and acquire seed URLs by sending them to a standard web search engine.

In the second step, we extend a standard crawler[2] by the text categorization tool TextCat ([3]), which creates classification models from training corpora by analyzing the frequencies of their character n-grams. Language specific models are distributed with the software. To create domain specific models we used two different document collections for each language. For each webpage fetched by the crawler we first test whether it is written in the required language. If so, we use the domain models in order to check the webpage's topic. Only those pages that match both language and domain are preserved.

[1] E.g. the crawler implemented in Nutch, see [7] for details.

[2] http://lucene.apache.org/nutch/

3. Experiments

In our experiments we crawled for medical webpages, separately for English and German. To distinguish between relevant and non-relevant documents we used clinical reference manuals and and newspaper articles acquired from the Internet. The crawls were started with 100 domain-specific URL seeds and terminated after the pre-defined depth of three links, followed starting from a seed URL, was achieved. In total, 9,850 webpages were downloaded for the English and 17,850 for the German scenario. We assume that the crawler accepted more German webpages, because its language model for the medical domain is less restrictive than the English one.

4. Results

The crawler roughly harvests 6 GByte text per day. In order to assess the quality of the corpora, two subjects evaluated manually two sample sets each, consisting of 200 documents randomly extracted from the English and the German crawled collection. Precision values, computed as the average percentage of the webpages satisfying the language (*P(L)*) and the topic (*P(T)*) condition, are shown in Table 1. The inter-rater agreement is perfect for the language scenario (1.0), but there is only a moderate agreement for judging the topic relatedness (0.5 for English and 0.6 for German medical texts) according to the $Kappa$[3].

Our crawler is highly language specific. Especially, for the German language, the performance of nearly 100% is very satisfactory, since the majority of the Internet sites is written in English. The evaluation of the topic relatedness shows that content-based focused crawling clearly outperforms crawler that rely purely on the link structure (e.g. Chakrabarti ([4]) reports harvest rate of about 40-45% by using automatic classification). It is also not enough to look only on specific keywords that appear in a document, cf. Stamatakis et al. ([10]), who achieved 92.1% for collecting English webpages related to the laptop domain (manual evaluation of a 150 sample). There is a great difference between the topic ratings for both languages. The German crawl data contains on average only 84% relevant webpages, while the English sample has an average precision of 97%. At the same time, webpages that were judged as unrelated in both scenarios mostly belong to related domains (pharmacy, biology, genetics etc.). As we supposed, the German classification models are not restrictive enough and need to be adjusted.

Scenario		Rater 1 (%)	Rater 2 (%)	Kappa
English	P(L)	99.5	99.5	1
crawl	P(T)	97.0	97.0	0.5
German	P(L)	99.5	99.5	1
crawl	P(T)	87.5	80.0	0.6

Table 1. Ratings of crawled webpages by their language and topic

[3] We compute $Kappa$ as $\mathcal{K} = \frac{P(A)-P(E)}{T-P(E)}$, where $P(A)$ is the observed agreement, T is the total number of examples and $P(E)$ is the agreement by chance.

5. Discussion

We do not strive to crawl all webpages related to medicine that are available on the Web, since it is unrealistic in terms of storage and crawling time. Therefore, we do not provide any recall values. Our main purpose is to have a large medical corpus, precise with regard to the domain and language focus, representative in terms of medical subdomains, but at the same time not too focused on any of them. As a complete download of entire web sites is of minor importance, we tolerate document ”gaps” as long as they do not result in a general bias of the whole corpus. Our further directions include large scale crawls and their evaluation, narrow classification of the crawled data and exploring its usefulness for multiple statistics-based NLP tasks.

References

[1] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In *Proceedings of the World Wide Web Conference*, pages 96–105, 2001.

[2] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Meeting of the Association for Computational Linguistics*, pages 26–33, 2001.

[3] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of the SDAIR’94, the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, Nevada, U.S.A, 1994.

[4] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1623–1640, 1999.

[5] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, and Marco Gori. Focused crawling using context graphs. In *26th International Conference on Very Large Databases, VLDB 2000*, pages 527–534, Cairo, Egypt, 10–14 September 2000.

[6] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web question answering: Is more always better? In *Proceedings of SIGIR 2002*, pages 291–298, 2002.

[7] R. Khare, D. Cutting, K. Sitaker, and A. Rifkin. Nutch: A flexible and scalable open-source web search engine. Technical report, CommerceNet Labs, 2004.

[8] P. Nakov and M. Hearst. A study of using search engine page hits as a proxy for n-gram frequencies. In *Proceedings of the RANLP’05*, 2005.

[9] K. Somboonviwat, T. Tamura, and M. Kitsuregawa. Simulation study of language specific web crawling. In *Proceedings of the SWOD’05*, 2005.

[10] K. Stamatakis, V. Karkaletsis, G. Paliouras, J. Horlock, C. Grover, J. R. Curran, and S. Dingare. Domain-specific web site identification: The crossmarc focused web crawler. In *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA 2003)*, pages 75–78, Edinburgh, UK, 2003.