Language Specific and Topic Focused Web Crawling

Olena Medelyan[1], Stefan Schulz[2], Jan Paetzold[2], Michael Poprat[2,3], Kornél Markó[2,3]

[1]University of Waikato, Department of Computer Science, Hamilton, New Zealand
[2]Freiburg University Hospital, Department of Medical Informatics, Freiburg, Germany
[3] Jena University, Computational Linguistics Research Group, Jena, Germany

Abstract

We describe an experiment on collecting large language and topic specific corpora automatically by using a focused Web crawler. Our crawler combines efficient crawling techniques with a common text classification tool. Given a sample corpus of medical documents, we automatically extract query phrases and then acquire seed URLs with a standard search engine. Starting from these seed URLs, the crawler builds a new large collection consisting only of documents that satisfy both the language and the topic model. The manual analysis of acquired English and German medicine corpora reveals the high accuracy of the crawler. However, there are significant differences between both languages.

1. Introduction

For most corpus based approaches to NLP document collections in a specific language that cover a particular domain are required. These collections are used as training data to build accurate statistical models reflecting their characteristics, constrained by the language, theme and style. Given these models unseen documents can be processed and analyzed to detect their language and category, create summary abstracts or extract new information not covered in the initial corpus. Large document collections are also important for linguists, who develop grammatical theories by analyzing our language.

Usually, large corpora are created manually and distributed to researchers through such organizations as Linguistic Data Consortium[1]. Professional linguists build document collections according to their language, domain, genre and type, which is a time-consuming and expensive task, especially when the corpora needs to be annotated, i.e. when part-of-speech tags, parse trees or word senses need to be assigned. Therefore, unsupervised approaches that use raw document collections become more popular among the NLP researchers. However, they still depend on the quality of the analyzed documents.

The obvious alternative to overcome tedious corpora acquisition is to use the Web instead, which is a mine of valuable language data. It has been successfully explored as training and test corpus for a variety of NLP tasks (Nakov and Hearst, 2005; Banko and Brill, 2001; Dumais et al., 2002). However, corpora derived from the Web are usually inconsistent and highly heterogeneous in their nature, which is normally counterbalanced by extending their size to billions of words. At the same time, statistics acquired from such large general corpora are imprecise, since a heterogeneous corpus contains a lot more ambiguities than a restricted one. The problem of detecting domain specific language characteristics also remains: domain-specific relations between words cannot be detected, they get lost among common word senses.

We assume that this problem can be solved by using *focused web crawling*. Given an initial small collection in a required language and domain, we propose to use a standard text classification approach combined with a crawling algorithm to acquire huge high-quality corpora from the Web. After an overview of related works on corpus acquisition and focused crawling, we describe our approach in detail. In our current experiments we have acquired large corpora with medical documents in English and German languages and used manual evaluation of sample sets from both collections to validate the accuracy of our approach. After presenting the evaluation results we discuss the encountered problems and our ideas for future experiments.

2. Related Works

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[2]. *Focused* crawling implies fetching only those pages that are relevant to a particular topic or language. Different approaches were proposed for the focusing strategy. In general, they all are based on the assumption that webpages on a given topic are more likely to link to those on the same topic. Starting from a set of webpages that represent the given topic, the crawler follows their links and is restricted by either content words on the outgoing webpages, or the graph structure of the Web, or URL tokens, or the combination of these criteria. They are used to prevent crawling of unrelated websites which would result in a deviation from the specific topic.

The focused crawler described by Chakrabarti et al. (1999) relies on model representing the linkage structure of the web, i.e. if a webpage points to a topic specific page, then other pages it is pointing to are likely to belong to the same topic. As an additional focusing restriction they use probabilistic classifiers built from a set of representative pages. The analysis of the graph structure presented in this work was adapted in various later studies on focused crawling (Diligenti et al., 2000; Aggarwal et al., 2001; Somboonviwat et al., 2005).

Diligenti et al. (2000) build context graphs from manually evaluated set of webpages, where seed webpages are linked

[1] `http://www.ldc.upenn.edu/`

[2] E.g. the crawler implemented in Nutch, see (Khare et al., 2004) for details.

to their parents and silblings. This structure is analyzed to extract hierarchical topic dependencies. Given these hierarchical classifiers Diligenti et al. (2000) designed a *context focused* crawler that outperforms the standard focused crawler in terms of required steps till a topic specific webpage is found. However, the authors admit that their approach depends on the nature of a category. It is only useful for categories that have a standard way of hierarchical positioning on the web, e.g. for finding papers on neuronal networks by using university websites.

Aggarwal et al. (2001) try to overcome the problem of category dependent link structures by combining them with other features, such as content words appearing on webpages, URL tokens etc. Depending on pre-defined arbitrary predicates, e.g. keyphrase sets or category names, their algorithm learns what features are more likely to guide the focused crawling.

These authors have in common that they assume the existence of similar graph structures existing on the Web. This is certainly the case for large web directories, such as Yahoo!, Ebay, or Amazon[3], however the rest of the web users do not construct web sites in a special structured way, simply because there are no widely accepted guidelines for this. At the same time, current attempts in focused crawling do not pay enough attention to the actual content of their training and test collections. They briefly mention considering content words appearing on the webpages as topic filters. Some use $TF \times IDF$ weightening schemes to restrict these sets to most representative words and speed up the classification process (Diligenti et al., 2000; Nekrestyanov et al., 1999). The selection of seed URLs is not described at all, and it is not clear how the seeds are related to the training data used in each case.

Ghani et al. (2001), in contrast, pay extensive attention to collecting focused seed queries in their approach. They use queries to acquire language specific corpora for minority languages such as Slovenian and Tagalog. The queries are automatically constructed from terms with the highest probability scores, computed from two sets with relevant and non-relevant documents. New documents retrieved with these queries are then additionally categorized with the text classification tool TextCat[4] to increase these initial sets. The results are encouraging: for example, out of 1000 retrieved documents, 835 were identified as correct. The disadvantage of this approach is the need of large contrast corpus and, most important, the absence of the crawling element. However, the experiments show the usefulness of TextCat for the corpus acquisition task.

We conjecture that selecting good starting points for the crawling process and the content analysis of the fetched webpages are important parameters in focused crawling, whereas graph structure can be ignored due to its inconsistency, as shown above. Given a sample collection of documents, we first select terms and phrases that represent its topic and create an extensive set of seed URLs using these phrases as queries. The seeds are starting points for the crawler, which is kept focused by TextCat, used both for language and topic specific categorization. The following section describes our approach in detail.

[3] Aggarwal et al. (2001) are using these websites as starting points for their crawls. This is rather an attempt of extracting a part of already manually categorized data, or its extension to linked webpages.

[4] `http://odur.let.rug.nl/~vannoord/TextCat`

3. Content-Based Focused Crawling

Our focused crawling is performed in two steps. Firstly, we create a list with topic and language specific seed URLs. Secondly, we run the open-source crawler Nutch[5] starting from these URLs and use the text categorization tool TextCat to avoid crawling of irrelevant webpages.

3.1. Collecting Topic-Specific URL Seeds

In order to collect seed URLs for the initialization of the crawling process, we first need domain specific queries. Given a document collection, we use a two to five words window to extract all possible phrases consisting of non-stopwords that appear in these documents. For each phrase i in our sample collection C we compute its average $TF \times IDF_{i,C}$ value, according to the following formula:

$$TF \times IDF_{i,C} = \frac{\sum_{c \in C} ((1+logTF_{i,c}) log \frac{|C|}{DF_i})}{|C|},$$

where $TF_{i,c}$ is term frequency, i.e. the frequency of a phrase i in the document c; DF_i is document frequency, i.e. the number of documents in the collection that contain the phrase i; and $|C|$ is the number of documents in the collection. Unlike Ghani et al. (2001) we do not need a contrast collection to distinguish between relevant and irrelevant phrases at this stage.

To ensure the topical relevance of each phrase, there might be an additional check, if it appears in a domain specific thesaurus. However, this step is not crucial, since top phrases ranked according their $TF \times IDF$ value, are already domain specific.

We use the resulting phrases as focused queries and acquire seed URLs by sending these queries to a standard web search engine.

3.2. Combining the Crawler with a Topic Filter

TextCat creates classification models from training corpora by analyzing the frequencies of their character n-grams (see Cavnar and Trenkle (1994) for details). Language specific models are distributed with the software. To create domain specific models we used two different document collections for each language. First, using all documents of our sample collection. Second, with a document collection from a different domain, which in our case consisted of news articles downloaded from the Internet.

Each time a webpage is fetched by the crawler, TextCat is used twice to ensure its similarity to the sample collection:

1. Does the language of the website corresponds to the language of the sample collection?
2. Does the topic of the website corresponds to the topic of the sample collection?

The webpage is only preserved if it passes both tests.

[5] `http://lucene.apache.org/nutch/`

4. Experiments and Crawling Statistics

Our experiments were conducted on a medical domain for two languages, English and German. The sample collections were medical articles downloaded from the Internet[6]. To create a contrast model for the focused filtering we used newspaper articles[7]. We prior extracted all unique paragraphs appearing in these collection, since repetitions are frequent in pages from the same website, and they influence frequency statistics. We have generated 100 queries for each language scenario and checked them against the medical thesaurus UMLS (UMLS, 2004). To increase the matching probability we remove stopwords from the UMLS terms as well and order words both phrases alphabetically.

In the first experiment, we used one top ranked URL per query and started the crawler with 100 seeds. It has 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.

In the second experiment, an extensive crawl was conducted, starting with 10 top ranked URLs per query and the crawling depth of 10 links. Given this data we analyzed how the *harvest rate*, i.e. the percentage of webpages satisfying the topical classification model, changes with the increasing depth, crawling time and the number of fetched webpages. Similar to other focused crawling approaches, the harvest rate increases quickly with the first crawled webpages and remains stable over in each case, after the first 15,000 pages are fetched. The crawler then accepts about 49% of English, and 58% of German webpages according to TextCat. Of course, these numbers depend on the quality of the training material and the restrictiveness of the models. Other focused crawler have been reported to have lower harvest rates (e.g. 40-45% in Chakrabarti et al. (1999) and 33-41% in Aggarwal et al. (2001)).

The third experiment was conducted with the same set of seed URLs, but the original version of the crawler, without the focusing element. Compared to our focused crawler, the harvest rate decreases quickly with the first crawled webpages, but it also stays stable after a certain amount of webpages has been fetched. In our scenario, the crawler fetched about 15% of medical webpages. This number might roughly reflect the amount of medical pages existing on the Web.

Although the speed of the crawler is affected by the TextCat classification, it still harvests reasonably fast, roughly 6 GByte text per day.

5. Manual Evaluation

In order to assess the quality of the corpora, we have randomly extracted separate sample sets from English and German crawled collections. Two subjects manually evaluated both of these sets, which consisted of 200 webpages each. The subjects had to answer whether the language and the topic of the given webpage was correct or not. If a webpage contained multiple languages, e.g. an English abstract and a German text, subjects were asked to consider the language used for the majority paragraphs. While language is easy to identify, assessment of topics has a certain amount of subjectivity. Therefore, subjects received a guideline for the topic evaluation: If a given webpage would be better understood by a health professional, then it is related to the medical domain.

Table 1 summarizes the results of the manual evaluation. We define precision as the average percentage of webpages which satisfy the language (*P(L)*) and the topic (*P(T)*) conditions. We use $Kappa$ to evaluate the inter-rater agreement in our experiment (Carletta, 1996):

$$\mathcal{K} = \frac{P(A)-P(E)}{T-P(E)},$$

where $P(A)$ is the observed agreement, T is the total number of examples and $P(E)$ is the agreement by chance. The results confirm that topic assessments were highly subjective. While subjects achieved perfect agreement for the language scenario (1.0), there was only a moderate agreement for judging the topic relatedness (0.5 for English and 0.6 for German medical texts).

The evaluation shows that our crawler is highly language specific. This is particularly true for the German language, for which we achieved the performance of nearly 100%, despite English being the predominant language of the Internet. The evaluation of the topic relatedness shows that, in the English scenario, our focused crawler outperforms other crawlers. For example, Stamatakis et al. (2003) achieved precision of 92% for collecting English webpages related to the laptop domain (manual evaluation of a 150 sample). The precision of our crawler is 5 percentage points higher (97%, cf. Table 1) on an even larger sample set.

There is a great difference between the topic ratings for both languages. The German crawl data contains on average only 84% relevant webpages, which is 13 percentage points lower than in the English sample. At the same time, webpages that were judged as unrelated in both scenarios mostly belong to related domains (pharmacy, biology, genetics etc.). The reason might be, as supposed, that the German classification models are not restrictive enough. Further investigations are required to find out the differences in language models and to adjust the performance of the crawler.

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

[6] `http://www.merck.com/mrkshared/mmanual/home.jsp` and `http://www.msd.de/msdmanual/home.html`

[7] `http://www.guardian.co.uk/` and `http://www.tagesspiegel.de/`

6. Discussion

The presented approach for focused crawling is simple, language independent and stable. We have extended a standard crawler by the text classification tool TextCat and started the crawling process with a few topic specific queries acquired with a simple statistical computation. Preliminary evaluations show that, in terms of precision, our focusing strategy outperforms other more sophisticated techniques based on computationally expensive graph analysis. However, manual evaluation of large-scale crawls is necessary, and the differences in the performance of English and German crawling need to be explained.

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 overly focused on any of them. We would tolerate document "gaps" as long as they do not result in a general bias of the whole corpus. One of our future goals is to find an appropriate evaluation of how well the crawled corpus represents subdomains of a topic compared to a manually created one. A comparison of how concepts are statistically distributed among both corpora would give valuable insights into this problem.

Our further directions include large scale crawls and their evaluation, extending our experiments to other languages and domains, narrow classification of the crawled data and exploring its usefulness for multiple statistics-based NLP tasks.

7. References

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

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

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

W. B. Cavnar and J. M. Trenkle. 1994. 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.

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

M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. 2000. Focused crawling using context graphs. In *26th International Conference on Very Large Databases, VLDB'00*, pages 527–534, Cairo, Egypt, 10–14 September.

S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. 2002. Web question answering: Is more always better? In *Proceedings of SIGIR'02*, pages 291–298.

R. Ghani, R. Jones, and D. Mladenic. 2001. Mining the web to create minority language corpora. In *Proceedings of the CIKM'01*, pages 279–286.

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

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

I. Nekrestyanov, T. O'Meara, A. Patel, and E. Romanova. 1999. Building topic-specific collections with intelligent agents. *Lecture Notes in Computer Science*, 1597:70–82.

UMLS. 2004. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

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

K. Stamatakis, V. Karkaletsis, G. Paliouras, J. Horlock, C. Grover, J. R. Curran, and S. Dingare. 2003. Domain-specific web site identification: The crossmarc focused web crawler. In *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA'03)*, pages 75–78, Edinburgh, UK.