Focused Crawling for Building Web Comment Corpora

Melanie Neunerdt, Markus Niermann and Rudolf Mathar
Institute for Theoretical Information Technology
RWTH Aachen University
Email: {neunerdt, niermann, mathar}@ti.rwth-aachen.de

Bianka Trevisan
Textlinguistics/Technical Communications
RWTH Aachen University
Email: trevisan@humtec.rwth-aachen.de

Abstract—**Web 2.0 provides various types of social media applications, e.g., blogs, forums and news sites that allow users to post Web comments. This kind of communication plays an important role in acceptance research. To extract different opinions from such data, it is necessary to build Web comment corpora. Building such corpora requires focused crawling. Many focused Web crawling algorithms are known to build topic-specific Web collections. However, the type of Web pages is typically not considered. In this paper, we introduce a new type-specific focused crawler, which uses a classifier based on HTML meta information. Its application allows for collecting only Web pages that cover Web comments from various domains.**

I. Introduction

Web 2.0 technologies allow users to participate and collaborate in creating and modifying the World Wide Web. The combination of technological progress and social drivers gives rise to constantly growing *user generated content* in the Web. This kind of communication is beneficent for acceptance research. Particularly, Web comments posted in blogs, forums, and news sites are very useful for extracting user opinions about a particular topic. Extracting opinions from Web comments serves as additional method for complementing traditional methods like questionnaires or interviews. In terms of acceptance studies, it is very important to study opinions from proponents, opponents as well as from impartial persons. Web comments from different communities are mostly located on different Web domains, therefore it is very important to obtain a Web comment corpus from many different domains. Building such Web comment corpora claims for the application of Web search techniques. Much research on different tools to build Web collections related to a specific topic has been done. The most popular approach is focused crawling [1].

Focused crawlers are designed to selectively seek out pages according to pre-defined relevance criteria. Generally, focused crawling algorithms aim at building a topic-specific Web collection without considering the type of the fetched Web page. The result is a collection of Web pages all dealing with the same topic, but without further restriction for the content type of the Web page. To obtain a collection of, e.g., Web pages to perform opinion detection for acceptance research, further data refinement has to be performed.

In this work, we develop a focused crawling algorithm which focusses on building Web comment collections. Therefore, a type-specific focused algorithm is proposed. Furthermore, we develop a classifier to predict relevance of a Web page with respect to the focused type. Topic-specifity is given by accurate seed page selection. Results are studied with respect to their type and topic relevance.

The outline of this paper is as follows: Section II summarizes the related work. In Section III we propose our focussing algorithm. Section IV and V describe our implementation and present some experimental results. Section VI covers the conclusion and discusses future work.

II. Related Work

In contrast to general search engines, focused crawlers selectively seek out Web pages related to a specific topic. According to a computed relevance score, target URLs are ordered in a priority queue and downloaded subsequently. This kind of search algorithm is called *best-first search* and is used in many focused crawlers [2], [3]. In [4] it is particularly pointed out that using *best-first search* the crawler could miss many relevant pages due to its local search property. Therefore, in [4] a new meta-search enhanced focused crawling approach is proposed to address the problem of local search. The task of type-specific Web search focussing on Web comments is strongly related to blog search algorithms. Several search and discovery tools for blogs, e.g., Blogdigger, Blogpulse and GoogleBlogsearch are available for accessing blog domains. In [5], capabilities and limitations of different blog search engines are analyzed and compared.

In addition to the Web search algorithm, an algorithm to determine the relevance of target URLs is needed. Many different algorithms have been proposed to solve the task of relevance scoring. Basically, two types are distinguished: *content-based relevance algorithms* and *link-based algorithms*. Content-based algorithms determine Web page relevance analyzing the actual HMTL content itself, whereas link-based algorithms analyze link structure between known Web pages to determine a relevance score. Common content-based approaches are presented in [1], [6], [3]. The most popular link-based approaches are Page Rank [7], HITS [8], and OPIC [9]. All three propose to calculate the relevance score on the link graph induced by pages fetched so far.

III. Focused Crawling Algorithm

To build German Web comment collections for opinion detection, we propose the following type-specific focused crawling approach. We assume all Web pages containing posted Web comments, to be relevant types. Furthermore, we assume that relevant pages are rather pointing to relevant pages than to irrelevant ones. Therefore, the algorithm rates target URLs located on relevant pages with higher scores. Two different algorithms are developed and evaluated: First, an algorithm with a soft focussing rule considering link- and content information to determine relevance scores, and second, a hard focussing rule considering only the content in relevance scoring.

To describe the algorithms we consider $k \in \mathbb{N}_0$ to be the current processed layer. In other words, k indicates the layer the crawler is currently working on. Layer 0 indicates the seed page layer and is increasing with growing search depth. $\mathcal{P}_k$ is the set of known Web pages at layer k. $\mathcal{P}_k$ is monotonically increasing with k and $\mathcal{P}_k \subseteq \mathcal{P}_{k+1}$. We define $\mathcal{M}_k \subseteq \mathcal{P}_k$ as the set of Web pages that have been fetched in layer k. Furthermore, we define a transition function $t_k(i,j) \in \mathbb{N}_0$ with $i, j \in \mathcal{P}_k$ which gives the number of links from i to j, since a Web page may contain more than one link to the same child page. Hence, the number of outgoing links $n_k(i)$ from page i can be calculated as

$$n_k(i) = \sum_{j \in \mathcal{P}_k \setminus \{i\}} t_k(i,j).$$

The first approach is a combination of a link-based and a content-based focussing algorithm: A relevance classifier is combined with the link-based *Online Page Importance Computation* (OPIC) scoring algorithm [9]. The OPIC algorithm works with a cash function $c_k(i)$ to compute Web page relevance. To combine the scoring with a relevance classifier, this cash function is adapted. First, we focus on the description of the adapted cash function. At the beginning of the algorithm, a constant amount of cash C is distributed equally between the seed pages at layer 0

$$c_0(i) = \frac{C}{|M_0|} \text{ , } \forall i \in M_0 \text{ .} \tag{1}$$

First, a relevance indicator function

$$\mathbf{1}_i = \begin{cases} 1 & \text{if page } i \text{ is classified as relevant} \\ 0 & \text{else} \end{cases} \tag{2}$$

is introduced, it calculates the relevance of a Web page as binary decision, i.e., a page can be relevant or non-relevant. Content-relevance of a Web page is predicted by means of a classifier which is described at the end of this chapter. Furthermore, we introduce the cash updating function for page j, which is performed at each iteration step of the OPIC algorithm as

$$c_{k+1}(j) = \sum_{i \in \mathcal{M}_k} \frac{t_k(i,j)}{n_k(i)} c_k(i) \text{ .} \tag{3}$$

For each fetched page $i \in \mathcal{M}_k$ the score is distributed proportionally to the number of links to child page i between child pages. To combine the OPIC algorithm with content-relevance feedback the cash function, (3) is modified to

$$c_{k+1}(j) = \sum_{i \in \mathcal{M}_k} \frac{t_k(i,j)}{n_k(i)} c_k(i) \mathbf{1}_i \text{ .}$$

Obviously, the algorithm follows the same cash distribution to child pages then OPIC for relevant pages in layer k. Non-relevant pages do not distribute any cash to child pages. To keep the total sum of cash constant at any time, the remaining undistributed cash from non-relevant pages in every layer

$$R_k = \sum_{i \in \mathcal{M}_k, \mathbf{1}_i = 0} c_k(i)$$

is equally distributed over relevant pages. Hence, we obtain

$$R_{k+1} = \frac{R_k}{\sum_{i \in \mathcal{M}_k} \mathbf{1}_i} \tag{4}$$

as additional cash for each relevant page in layer k. Distributing the additional cash between all relevant pages, equation (III) becomes

$$c_{k+1}(j) = \sum_{i \in \mathcal{M}_k} \frac{t_k(i,j)}{n_k(i)} \left(c_k(i) + R_{k+1} \right) \mathbf{1}_i \tag{5}$$

as new cash updating function. Note that the cash sum in each layer is equal to C. The total relevance of a Web page $j \in \mathcal{P}_{k+1}$ after each iteration, is defined as score

$$s_{k+1}(j) = \sum_{n=0}^{k+1} \frac{c_n(j)}{C} \text{ , i.e.,}$$

as sum of relative cash weights over all evaluated layers. According to the scores $s_{k+1}(j)$ all Web pages $j \in \mathcal{P}_{k+1}$ are ordered in the priority queue and downloaded subsequently.

Furthermore, a second focussing algorithm is developed. According to the content-relevance of a Web page, a binary decision is made for further searching, without respect to the link-structure of Web pages. Using the indicator function from equation (2) the score

$$s_{k+1}(j) = \begin{cases} 1 & \sum_{i \in \mathcal{M}_k} \frac{t_k(i,j)}{n_k(i)} \mathbf{1}_i > 0 \\ 0 & \text{else} \end{cases}$$

is defined, i.e., Web page scores of child pages only depend on the relevance of their direct parent Web pages. At least one parent page has to be predicted as relevant. For this algorithm a binary decision is made, therefore Web pages in the priority queue can only be ordered according to 0 or 1 scorings. In this case we randomly select relevant Web pages with score 1 to be fetched in the next layer.

For our purposes, relevance prediction for Web pages requires the detection of Web comments. Therefore, a two-class problem has to be solved. A Web page may be assigned to the class *comment* if it contains at least one Web comment or to the class *non-comment* elsewise. Properties of the HTML syntax, i.e., *Cascading Style Sheets* (CSS) classnames are selected as classification features. According to a given keyword list a binary decision is made: If a Web page contains at least one CSS classname, where one of the keywords is a substring, the Web page is classified as *comment* and as *non-comment* if no keyword is detected. To compose a list of common keywords, CSS classnames located closely to Web comments are collected from a set of training data. The training data set comprises Web pages from different domains to receive a representative list. In total, a list of 25 keywords is composed.

IV. Implementation

We compare our approaches *OPIC combined with comment detection* (OPIC+COMMD) and *comment detection* (COMMD) focus algorithm to the standard *breath-first* (BFIRST) search crawling approach, [4]. The BFIRST algorithm does not use any heuristics when deciding which URL to visit next and downloads Web pages successively in the order they are discovered.

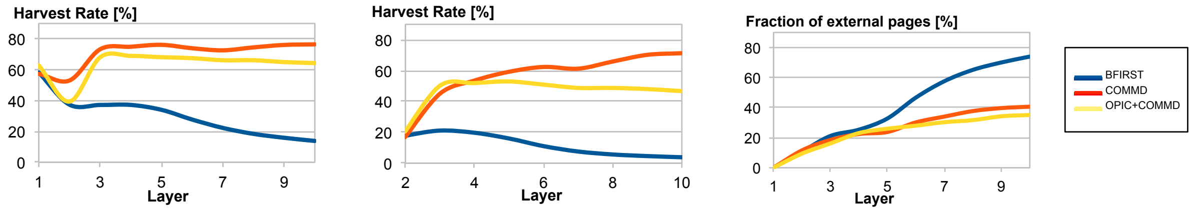
In order to examine the performance of our proposed algorithms, the open-source search engine Nutch, [10], is extended. In our approach, a new plugin is developed to realize the type-specific focus algorithm. The Nutch search engine allows for some additional parameter settings. Particularly, the search *depth* and the maximum number of pages fetched at one layer *topN* serve as stopping criteria for the algorithm. If on any layer there are more URLs to be downloaded, the crawler dismisses those ones with the lowest page rank. For our purposes, we choose *topN*=10000 and *depth*=10 for all three crawling processes. In addition to our scoring filter function, an URL filter is activated to disable some popular sites like *Skype*, *Ebay* and *Amazon*, that are highly interconnected but obviously irrelevant for our investigations. Initially, seed pages are selected according to topics that are related. Exemplarily, topics are chosen according to our project dealing with *acceptance aspects in mobile communications*. The selected Web pages either contain discussions fitting the topic or hub Web sites containing important links to URLs related to that topic. In total 112 seed pages from 78 different domains are selected.

V. Experimental Evaluation

For the evaluation of our crawling algorithms, we consider the harvest rate. The harvest rate represents the fraction of relevant Web pages, divided by the set of all fetched Web pages. First, harvest rates for all three algorithms are depicted in figure 1 (a). The plot shows the harvest rates over increasing layer k. With increasing layer depth, harvest rates for BFIRST decrease continuously to 14%. Both presented algorithms considerably exceed the harvest rates of BFIRST algorithm for each layer k. Furthermore, up to layer 10 rates increase to 64% applying the OPIC+COMMD algorithm and up to 76% applying COMMD algorithm. Generally, it is important to detect Web comments particularly on new domains, i.e., not located on seed page domains. Therefore, we determine the fraction of *external* pages, i.e., pages that are not located on seed page domains, for each crawling process. Plot 1 (c) shows the result per layer. BFIRST algorithm yields the best result (76%) which can easily be explained by the fact that no relevance criteria have to be fulfilled for any page. OPIC+COMMD and COMMD algorithms end up between $30\% - 40\%$ which is still sufficient for our purposes. To investigate fetched external Web pages in more detail, we calculate harvest rates on these Web pages only. Results are depicted in 1 (b): The BFIRST algorithm leads to a high number of external pages, but harvest rates are significantly lower compared to our proposed algorithms. Particularly, the proposed COMMD algorithm shows a trade-off between good harvest rate and the fraction of fetched external pages.

The evaluations so far analyze the ability of the proposed algorithms to focus on specific text-types, i.e., Web comments. However, we are interested in the topic relevance of the collected Web pages. The same set of 1,000 Web pages later used for the classifiers evaluation is evaluated manually according to topic-relevance. Results prove that Web comments dealing with *acceptance aspects in mobile communications* are rare with 4.9%. However, 68% of the analyzed data are at least dealing with mobile communication systems in general. This result shows that the corpus is focused on an extended topic. Furthermore, we analyze the similarity of fetched pages between the three different focussing algorithms, to investigate the influence of different Web page relevance scoring on the resulting Web comment corpus. Figure 2 shows the results. On the one hand intersection sets of Web pages and on the other hand domain intersection sets are depicted. For a high corpus quality it is important to obtain Web comments from different domains and communities, such that opinions from different spectators are represented. Applying OPIC+COMMD results in 2,791 different domains, which exceeds the number of fetched domains considerably compared to BFIRST and COMMD algorithms, see figure 2. 84.5% of the domains appearing in the Web comment corpus build with COMMD are also appearing in the corpus build by using the OPIC+COMMD algorithm. These results approve that extending pure content-based scoring algorithms with link-based scoring, opens up the way to reach different domains in the Web and to obtain a better corpus quality, particularly in terms of opinion detection for acceptance research. Comparing the identical Web pages for all three algorithms shows that there is not much similarity in results. The highest similarity (around 32%) is observed comparing our proposed algorithms OPIC+COMMD and COMMD.

To calculate accurate harvest rates for the different focussing algorithms, it is necessary to know if a Web page is relevant or not. Our goal is to detect any kind of Web pages that include Web comments. Since a manually relevance evaluation of about 70,000 Web pages is very time consuming, a relevance metric is needed. In chapter III, we described the classifier which is used to focus on such kind of Web pages. At the same time, the classifier is used as relevance metric to estimate harvest rates. Some manual Web page classification (training



(a) Total pages.

(b) External pages.

(c) Fraction of external pages.

Figure 1. Harvest rate for total crawling results and only pages which are not located on seed page domains.

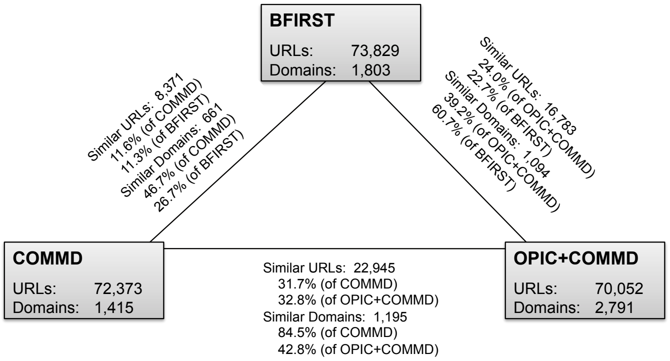


Figure 2. Similarity between crawling algorithms.

data) is performed to determine the classifiers accuracy and the relevance metric respectively. The classifier accuracy has to be considered in harvest rate estimates. Therefore, 1,000 sample Web pages are drawn randomly from the Web comment corpus derived from applying the COMMD algorithm. The samples are manually classified and compared to their predicted classes. The two class problem is evaluated, where a Web page is labeled as *comment* or *non-comment*. Table I depicts the confusion matrix results. In total a classification accuracy of 86% is achieved with our approach for the selected 1,000 samples. Furthermore, a precision rate of 0.86 and a recall rate of 0.88 can be observed for the *comment* class.

Table I
CONFUSION MATRIX OF THE TWO CLASS PROBLEM.

	Predicted class	
True class	Comment	Non-Comment
Comment	482	63
Non-Comment	77	378

VI. CONCLUSIONS AND FUTURE WORK

The developed algorithms allow for type-specific focused crawling to build Web comment corpora. Best harvest rates can be achieved by applying the proposed COMMD algorithm based on a content-based relevance scoring (classifier) for Web page ranking. Combining content-based scoring with link-based OPIC scorings, results in slightly lower harvest rates but increases the number of visited domains, which leads to a high Web comment corpus quality. Both proposed algorithms exceed harvest rates obtained by a standard breath-first search algorithm. Furthermore, a classifier is proposed which leads to high precision and recall rates for Web comments and achieves a classification accuracy of 86% for the purpose of type-specific focused crawling.

Future work will be carried out in several directions. To assure topic-specificity in the resulting Web comment corpus, topic classification will be integrated into the relevance scoring of Web pages. For further corpus refinement, an ontology-based corpus generation tool proposed in [11] will be integrated. Furthermore, we want to improve Web comment detection by a more sophisticated approach.

ACKNOWLEDGMENT

This work was partially supported by the Project House HumTec at RWTH Aachen University, Germany.

REFERENCES

[1] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific web resource discovery," in *Computer Networks*, 1999, pp. 1623–1640.

[2] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through url ordering," in *Seventh International World-Wide Web Conference*, 1998.

[3] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Building domain-specific search engines with machine learning techniques," in *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.*, 1999.

[4] J. Qin, Y. Zhou, and M. Chau, "Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method," in *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*.

[5] G. Mishne and M. de Rijke, "A study of blogsearch," in *28th European Conference on IR Research*.

[6] S. Chakrabarti, K. Punera, and M. Subramanyam, "Accelerated focused crawling through online relevance feedback," in *Proceedings of the 11th international conference on World Wide Web*.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Technical Report, November 1999.

[8] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.

[9] S. Abiteboul, M. Preda, and G. Cobena, "Adaptive on-line page importance computation," in *WWW*, 2003, pp. 280–290.

[10] R. Khare, D. Cutting, K. Sitaker, and A. Rifkin, "Nutch: A flexible and scalable open-source web search engine," Tech. Rep., 2004.

[11] M. Neunerdt, B. Trevisan, T. C. Teixeira, R. Mathar, and E.-M. Jakobs, "Ontology-based corpus generation for web comment analysis," in *Proceedings of The ACM Conference on Hypertext and hypermedia*, Eindhoven, May 2011.