Guide Focused Crawler Efficiently and Effectively Using On-line Topical Importance Estimation*

Ziyu Guan, Can Wang, Chun Chen, Jiajun Bu, and Junfeng Wang
College of Computer Science, Zhejiang University
Hangzhou 310027, China
{guanzh, wcan, chenc, bjj, wangjunfeng}@zju.edu.cn

ABSTRACT

Focused crawling is a critical technique for topical resource discovery on the Web. We propose a new frontier prioritizing algorithm, namely, the OTIE (On-line Topical Importance Estimation) algorithm, which efficiently and effectively combines link-based and content-based analysis to evaluate the priority of an uncrawled URL in the frontier. We then demonstrate OTIE's advantages over traditional prioritizing algorithms by real crawling experiments.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*

General Terms: Algorithms, experimentation.

Keywords: Focused crawlers, topical crawlers.

1. INTRODUCTION

Focused crawlers, also called topical crawlers, are designed for topical resource discovery from the Web. Most focused crawling algorithms are variations of the best-first-search (BFS) algorithm in which the crawl frontier is maintained as a priority queue. The key research issue is how to properly prioritize the crawl frontier in order to focus the crawler on a topic. The foundation work of focused crawling was done by Chakrabarti *et al.* [2].

Traditional focused crawlers exploit contents (typically with a classifier) in the downloaded pages (i.e. link contexts) to predict relevance of the unvisited URLs and prioritize the frontier accordingly (we call this strategy *Link-context-prediction*, e.g. [4]). However, content-based retrieval methods usually demonstrate a poor performance in the Web. This is partly due to the noisy contents in Web documents i.e. documents with little text, containing images, scripts and other types of data which cannot be used by content-based methods, but also partly because these documents are created by different authors, with no coherence in style or structure.

Another source of information that can be exploited in Web information retrieval (IR) is the link structure of the Web. Our intuition is to assess the quality of Web pages with respect to the concerned topic and rank the frontier accordingly. In the literature there are some work combining content and link analysis for focused crawling. However, they all need to store the Web graph crawled so far and apply the heuristic algorithm periodically to the data structure. As the crawling proceeds, the Web graph visited will grow continuously and so will the time needed to execute the algorithm and the memory needed to store the Web graph, which significantly decrease the performance of the crawler. We propose a new frontier prioritizing algorithm inspired from the On-line Page Importance Computation (OPIC) algorithm [1], namely On-line Topical Importance Estimation (OTIE). OTIE is scalable and takes both link and content evidences into account.

*The project is sponsored by S&T Planning Projects of Zhejiang Province, No.2007C23086.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. OTIE

The general idea of OTIE is similar to that of OPIC [1]. OPIC computes PageRank in an on-line fashion. Like in OPIC, we transfer "cash" (i.e. importance) among pages in OTIE. However, our method differs from OPIC in that we bias cash distribution in OTIE to favor on-topic pages and to suppress off-topic pages. Let O_p denote the set of pages page p points to. When a page i is crawled, we distribute its cash in accordance with the similarity scores of the pages in O_i with respect to the concerned topic. It means the cash a page j in O_i gains from page i is proportional to its similarity to the topic:

$$\forall j \in O_i,\ cash\_gain(j) = \frac{\mathrm{sim}'(j,t)}{\sum_{u \in O_i} \mathrm{sim}'(u,t)} \times cash(i) \quad (1)$$

where the similarity function $\mathrm{sim}'(j,t)$ is defined as:

$$\mathrm{sim}'(j,t) = \begin{cases} \mathrm{sim}(j,t) & j \in S_{\mathrm{fetched}} \\ \mathrm{predicted\_sim}(j,t) & j \in S_{\mathrm{unfetched}} \end{cases} \quad (2)$$

Here S_{fetched} denotes the set of pages the crawler has downloaded, while $S_{\mathrm{unfetched}}$ denotes the set of uncrawled pages in the frontier. For an uncrawled page we use its link context in currently fetched page to predict its similarity. However, using predicted similarities introduces the problem of inaccurateness: link contexts of an URL sometimes cannot predict the relevance of the corresponding page of that URL appropriately, and hence make the distribution of cash unfair. To remedy this situation, we introduce a revision phase into the algorithm: when a page is crawled, the cash value is firstly revised according to the formula:

$$\Delta cash = cash \times \max\left(-1, a(2r-1)^d\right) \quad (3)$$

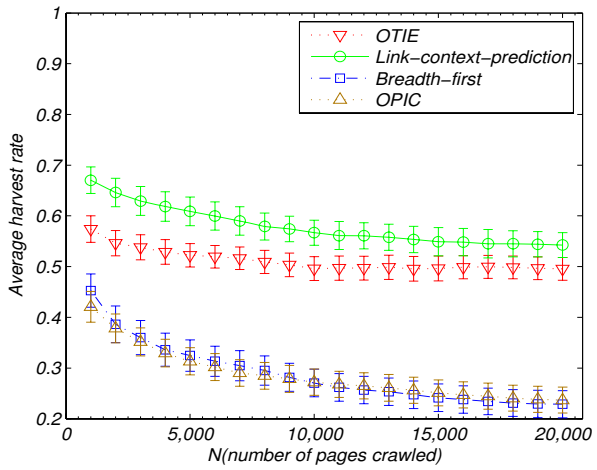

Figure 1: Harvest rate of the 4 crawling strategies averaged over 30 topics during the crawling process.

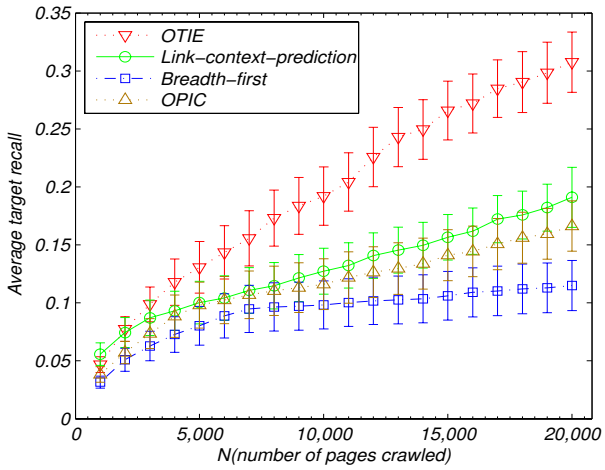

Figure 2: Target recall of the 4 crawling strategies averaged over 30 topics during the crawling process.

where r is the topical relevance score of the currently crawled page and a, d the parameters satisfying the following constraints: $a > 0, d \in \{d|d > 0$ and x^d is an odd function$\}$. Parameter d controls the flatness of the curve of function $a(2r-1)^d$ when r is near 0.5. Parameter a controls the absolute degree to which cash is altered.

The whole algorithm goes as follows: initially, the total amount of cash is equally distributed among the set of seed pages. Afterward in crawling, when a page is crawled, its cash is revised according to equation (3). Then according to equation (1), the cash is distributed among pages corresponding to the hyperlinks found on that page. The priority of an uncrawled page is the amount of cash it possesses. Periodically, the previously fetched page with the highest cash is recrawled to distribute its cash to further reward the pages it points to. The crawling process continues until a sufficient number of pages are fetched or there is no unvisited URL.

3. EXPERIMENTS AND RESULTS

Real crawling experiments were conducted over 30 topics selected from Open Directory Project (ODP)[1]. Three types of classifiers are used for content analysis, i.e Support Vector Machine (SVM), Naïve Bayes (NB) and Neural Network (NN). We use the SVM classifier to estimate (predicted) similarities of Web pages to the topic, and use all the three classifiers to classify Web pages. Web pages corresponding to the resource links of the 30 topics are used to train and test classifiers. Text contents extracted from Web pages are represented as vectors using the well-known Vector Space Model (VSM) with the TF-IDF weighting scheme. To evaluate focused crawling algorithms, we use two surrogate metrics *harvest rate* [3, 4] and *target recall* [3, 4] that approximate the two standard IR evaluation metrics *precision* and *recall* respectively. We evaluated four different crawling strategies: *Breadth-first*, *Link-context-prediction*, *OPIC* (*greedy* strategy) and *OTIE*. Breadth-first is treated as a baseline. OPIC is evaluated in order to confirm that "biasing" leads to a significant performance improvement. We adopt the following definition of link context to compute predicted similarities in Link-context-prediction and OTIE:

$$score = \beta * page\_score + (1-\beta) * context\_score \quad (4)$$

where *page_score* is the relevance score of the entire page content, *context_score* is the relevance score of a text window around a hyperlink on that page, and β is the relative weight assigned to *page_score*. We set $\beta = 0.25$ and text window size $T = 20$ words (including the anchor text). We also carried out preliminary experiments to investigate the proper parameter values for a and d in equation (3). We do not show the details of preliminary experiments because of the limitation of space. In this paper we set a=0.85 and d=3.0.

Figure 1 and Figure 2 show the experiment results. OTIE significantly outperforms the other strategies on average target recall. Link-context-prediction's precision is the highest. However, as more and more link evidence is seen, it seems that the precision of OTIE is approaching that of Link-context-prediction.

4. CONCLUSIONS

This paper proposed a new frontier prioritizing algorithm, OTIE, which requires neither storage of the Web graph crawled so far nor periodic application of a heuristic algorithm. It efficiently and effectively combines link-based and content-based analysis to evaluate the benefit of following an uncrawled URL.

5. REFERENCES

[1] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *Proceedings of the Twelfth International Conference on World Wide Web*, pages 280–290. ACM Press, 2003.

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

[3] G. Pant and P. Srinivasan. Learning to crawl: Comparing classification schemes. *ACM Trans. Information Systems*, 23(4), 2005.

[4] G. Pant and P. Srinivasan. Link contexts in classifier-guided topical crawlers. *IEEE Trans. Knowledge and Data Engineering*, 18(1), 2006.

[1]http://www.dmoz.org