News Page Discovery Policy for Instant Crawlers

Yong Wang, Yiqun Liu, Min Zhang, and Shaoping Ma

State Key Lab of Intelligent Tech. & Sys., Tsinghua University
wang-yong05@mails.tsinghua.edu.cn

Abstract. Many news pages which are of high freshness requirements are published on the internet every day. They should be downloaded immediately by instant crawlers. Otherwise, they will become outdated soon. In the past, instant crawlers only downloaded pages from a manually generated news website list. Bandwidth is wasted in downloading non-news pages because news websites do not publish news pages exclusively. In this paper, a novel approach is proposed to discover news pages. This approach includes seed selection and news URL prediction based on user behavior analysis. Empirical studies in a user access log for two months show that our approach outperforms the traditional approach in both precision and recall.

Key words: web log, user behavior analysis, news page discovery.

1 Introduction

Nowadays, there are high freshness requirements for search engines. Many web users prefer reading news from search engines. They type a few key words about a recent event into a search engine, check the returned result list and navigate to pages providing details about the event. If a search engine fails to perform such service, users will be frustrated and turn to other search engines. News pages should be downloaded immediately after they are published. Therefore, many search engines have special crawlers called instant crawlers to download novel news pages. The work flow of an instant crawler is

```
load seed URLs into waiting list                          (1)
while (waiting list is not empty)
{
pick a URL from the waiting list
download the page it points to
write the page to disk
for each URL extracted from the page
if the URL points to a novel news page                    (2)
add the URL to the waiting list
}
```

The performance of an instant crawler is largely determined by two factors: (1) quality of seed URLs; (2) accuracy of prediction about whether a URL points to a news page when its content has not been downloaded yet.

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 520–525, 2008.

Currently, manually generated rules are provided to solve the problem. An instant crawler administrator writes a news website list for an instant crawler to monitor. The instant crawler takes the homepages of these websites as seed URLs. A newly discovered URL will be added to its waiting list if it is in the monitored websites.

This policy works fine, but there are some problems. Many web sites contain both news pages and non-news ones. For example, *auto.sohu.com* is a website about automobiles. There are news pages reporting car price fluctuation and non-news pages providing car maintenance information. Only news pages in this website should be downloaded by instant crawlers. A web site is too large a granularity to make this discrimination. This problem can be solved with our method.

News pages provide information on recent events. Users are interested in a news page only in a short period after it is published. As more and more users get to know the event, fewer users are likely to read that page. In contrast, non-news pages are not relevant to recent events. Users access them constantly. This feature is used to identify news pages. If a page accumulates a large proportion of click throughs in a short period after publication, it is likely to be a news page.

A policy for instant crawlers to discover news pages is proposed based on user behavior analysis in click through data. In the beginning, news pages are identified based on how their daily click through data evolves. Then web pages which directly link to many news pages are used as seed URLs. Web administrators usually publish news pages under only a few paths, such as /news/. URLs of many news pages in the same folder share the same news URL prefixes. If there are already many news pages sharing the same news URL prefix, it is likely that novel news pages will be stored under that path and their URLs will start with that prefix.

The rest of this paper is organized as follows: Section 2 introduces earlier research in priority arrangement in waiting list of crawlers; Section 3 describes the dataset which will be used later; Section 4 discusses and verifies a few properties of news pages; Section 5 addresses the problems in seed selection and news URL estimation; the approach proposed is applied in the dataset and the result is analyzed in Section 6; Section 7 is the conclusion of this paper.

2 Related Work

Earlier researchers performed intensive studies on evolutionary properties of the web, including the rate of existing page updates and that of novel page appearance [1], [2]. The conclusion is that the web is growing explosively [2] and it is almost impossible to download all novel pages. Web crawlers have to organize a frontier which is consisted of discovered but not downloaded URLs. Priority arrangement in the frontier is important. This problem is studied from several perspectives. Some researchers tried to find a balance between downloading novel pages and refreshing existing pages [3], [4] and [5]. They studied page update intervals and checked existing pages only when necessary. Crawlers downloaded novel pages during the intervals. Focused crawlers only download pages related to a given topic [6], [7], [8] and [9]. They estimate whether a URL is worth downloading mainly based on its anchor text. Other crawlers [10], [11], [12] and [13] predict quality of novel URLs and download candidates of high quality. This work is similar with ours. We also make an order of the frontier, in the perspective of freshness requirements instead of page quality. Pages of high freshness requirement are downloaded with high priority, while others can be downloaded later.

3 News Page Discovery Policy for Instant Crawlers

News hub pages are used as seed URLs to discover novel news pages if they link to many previous news pages. Novel news pages are usually stored in the same location with known news pages. So news pages are identified to find where novel news pages are likely to be stored. A newly discovered URL will be downloaded if its URL starts with one of the news URL prefixes.

3.1 Generate Seed URL List for an Instant Crawler

It is proved in Section 4.1 that ClickThroughConcentration of most news pages is larger than that of most non-news ones. For each web page in the click through log, it is a news page if its ClickThroughConcentration is less than a threshold. Otherwise, it is a non-news page. News pages can be automatically identified with this method.

A seed URL for an instant crawler is of high quality if a large number of news pages can be discovered from it in only one or two hops. It is probable that novel news pages will be linked by pages which already have links to many known news pages. News hub pages which have linked most news pages are included in seed list.

3.2 Estimate Whether a URL Points to a News Page

Some news pages cluster in the same folder and some are dynamically generated from the same program with different parameter values. News URL prefixes can be found from known news pages. Given a website, a URL prefix tree is built according to its folder structure. In this tree, a node stands for a folder. Node A has a child node B if B is the direct subfolder of A. Web pages are leaf nodes. A program is also a non-leaf node. Dynamic pages generated from that program are its leaf nodes. Each non-leaf nodes are labeled by two numbers: the number of news pages and that of non-news

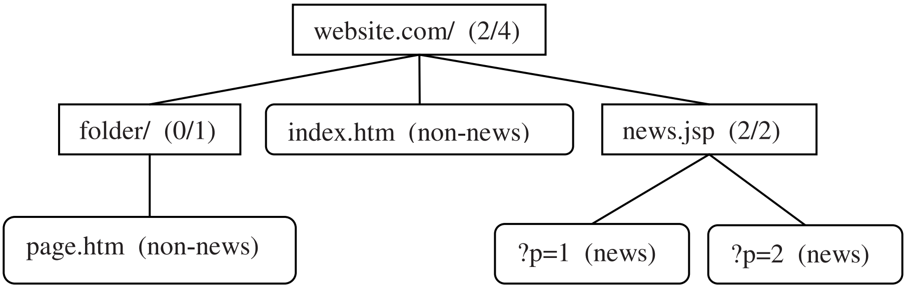

Fig. 1. A URL prefix tree of a sample website

ones directly and indirectly under that node. For example, website.com contains four pages: /index.html, /folder/page.htm, /news.jsp?p=1 and /news.jsp?p=2. Its URL prefix tree is organized as in Fig. 1.

Each non-leaf node is scored with the proportion of the number of news pages to that of all pages under that node. All prefix trees are traversed from the roots. A node is a news node if its score is greater than a threshold. Otherwise, its children nodes are tested. This algorithm is described below.

```
FindNewsNode(TreeNode N){
if (score of N is greater than the threshold){
N is a news node;
return;
}
foreach child in NonLeafChildrenOfN
FindNewsNode(child);
}
```

A news URL prefix consists of all nodes on the path from the root to the news node. Take the tree in Fig. 3 for example, if the node “news.jsp” is a news node, “website.com/news.jsp” is a news URL prefix. It is probable that URLs starting with news URL prefixes point to news pages and is worth downloading.

4 Experiment and Evaluation

Anonymous click through data for consecutive 60 days from November 13th 2006 to January 11th, 2007 is collected by a proxy server. Each record is a structure below:

Request Date and Time	Client IP	Target URL	Referrer URL

A user accesses the Target URL from a hyperlink in the Referrer URL. Referrer URL is null if a user types the address instead of clicking a hyperlink. Daily click through data of all 75,112,357 pages is calculated. Multiple requests to a single page from the same IP in one day are counted as one click through to avoid automatically generated requests by spammers. Pages whose average daily click throughs are less than one are filtered out for lack of reliability, leaving daily click through history of 975,151 pages for later studies.

4.1 Experiment

A page is classified as a news page if its ClickThroughConcentration is greater than a threshold p. Pages from focus.cn which have been annotated manually are used as training set in which there are 2,337 news pages and 703 non-news pages. The best performance is achieved when p=1.91 and the maximized hit (the number of pages correctly classified) is 2,682. This threshold is applied on all pages and 147,927 are labeled news pages and other 827,224 are labeled non-news.

A navigation record from page A to page B indicates that B is linked by A. The number of news pages linked by each page is calculated and the top 1,542 pages which link to the most news pages are included in the seed URL list. The number of seed URLs is the same with that in the baseline used later for comparison.

In URL prefix trees, a node is a news node if the proportion of news pages under that node is larger than a given threshold, where 0.8 is used. 439 nodes are labeled news nodes. Larger threshold can be used if bandwidth is limited and that wasted in downloading non-news pages is unaffordable. If an instant crawler has enough bandwidth and wants to recall more news pages, the threshold can be smaller.

4.2 Evaluation

Sogou Inc. is a search engine company in China. Its instant crawler uses a manually generated website list which contains 1,542 news websites. Homepages of these websites are seed URLs and the instant crawler downloads pages from these websites only. This policy is used as the baseline to be compared with ours.

Table 1. Performance comparison

	Baseline	Our Method
Number of Downloaded News Pages	86,714	101,870
Number of Total Downloaded Pages	177,801	111,934
Precision	48.8%	91.0%
Recall	58.6%	68.9%

46,210 different news pages are directly linked by homepages of news sites in Sogou's list, while 79,292 are directly linked by news hub pages in our seed list. Not all homepages are the best seeds. There are websites which publish both news pages and non-news ones. The index pages of news channels are better candidates for seed URLs. For example, finance.sina.com.cn is a financial website. The web log shows that most of its news pages are linked by finance.sina.com.cn/stock/. This page instead of the homepage should be included in the seed list.

The instant crawler of Sogou Inc. downloads all pages from their site list, while our instant crawler downloads pages whose URLs start with one of the news URL prefixes. The result is shown is Table 2.

There are 147,927 news pages in the dataset. Precision is the proportion of downloaded news pages in all downloaded pages. Recall is the proportion of downloaded news pages in all news pages in the data set. As is shown in Table 2, 86,714 news pages are in the site list, while 101,870 are covered by the URL prefixes. The performance of the efficiency crawler is improved that it downloads more news pages with less burden of non-news ones.

5 Conclusion

In this paper, an effective news page discovery policy is proposed. The current instant crawlers which are assigned to download news pages cannot produce satisfactory result due to news page distribution complexity. In this paper, we propose and verify a few features of news pages. Then these features are used in seed URL selection and news URL prediction. The performance of instant crawlers is improved both in precision and recall because they can discover more news pages with less bandwidth wasted in downloading non-news pages.

References

1. Fetterly, D., Manasse, M., Najork, M., Wiener, J.L.: A Large-scale Study of the Evolution of Web Pages. Software Practice and Experience (2004)
2. Brewington, B., Cybenko, G.: How Dynamic is the Web. In: Proceedings of WWW9 –9th International World Wide Web Conference (IW3C2), pp. 264–296 (2000)
3. Cho, J., Garcia-Molina, H.: Effective Page Refresh Policies for Web Crawlers. ACM Transactions on Database Systems (TODS) (2003)
4. Shkapenyuk, V., Suel, T.: Design and Implementation of a High-performance Distributed Web Crawler. In: Proceedings of the 18th International Conference on Data Engineering, San Jose, Calif. (2002)
5. Barbosa, L., Salgado, A.C., Carvalho, F., Robin, J., Freire, J.: Workshop On Web Information And Data Management. In: Proceedings of the 7th annual ACM international workshop on Web information and data management (2005)
6. Menczer, F., Belew, R.: Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web. Machine Learning 39(23), 203–242 (2000)
7. Pant, G., Menczer, F.: Topical Crawling for Business Intelligence. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 233–244. Springer, Heidelberg (2003)
8. Stamatakis, K., Karkaletsis, V., Paliouras, G., Horlock, J., et al.: Domain-specific Web Site Identification: the CROSSMARC focused Web crawler. In: Proceedings of the 2nd International Workshop on Web Document Analysis (WDA 2003), Edinburgh, UK (2003)
9. Menczer, F., Pant, G., Srinivasan, P.: Topical Web Crawlers: Evaluating Adaptive Algorithms. ACM Transactions on Internet Technology 4(4), 378–419 (2004)
10. Cho, J., Garcia-Molina, H., Page, L.: Effecient Crawling through URL Ordering. WWW8 / Computer Networks 30(1-7), 161–172 (1998)
11. Eiron, N., McCurley, K.S., Tomlin, J.A.: Ranking the Web Frontier. In: Proc. 13th WWW, pp. 309–318 (2004)
12. Eiron, N., McCurley, K.S.: Locality, Hierarchy, and Bidirectionality in the Web. In: Workshop on Algorithms and Models for the Web Graph, Budapest (2003)
13. Abiteboul, S., Preda, M., Cobena, G.: Adaptive On-line Page Importance Computation. In: Proc. 12th World Wide Web Conference, pp. 280–290 (2003)