News Page Discovery Policy for Instant Crawlers

Yong Wang, Yiqun Liu, Min Zhang, 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 requirement 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 Otherwise, they will become outdated soon.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 download 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 Bandwidth is wasted in downloading non-news pages because news websites do not publish news pages exclusively.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 on a user access log for two months show that our approach outperforms the traditional approach in both precision and recall.

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, while traditional common crawlers are assigned to download novel non-news pages and check updates of existing pages. The work flow of an instant crawler is

```
load seed URLs into waiting list                    (1)
while (waiting list is not empty)
{
    pick up 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.

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 starts from 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 web sites.

This policy works fine, but there are still some problems. First, news websites which are not so famous are likely to be omitted from seed list, even though they contain high quality news pages. Second, some web sites are associated with particular events. For example, doha2006.sohu.com reported the 15th Asian Games. It provided news service only during the 15th Asian Games session. The crawler administrator may not be so sensitive to the emergence and disappearance of temporary news websites. Third, there are many web sites which contain both news pages and non-news pages. For example, auto.sohu.com is a website about automobiles. There are news pages reporting car price fluctuation and also 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. These problems 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. Non-news pages are not relevant to recent events. Users are always interested in these pages and 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 probable 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, so instant crawlers can start with these seed URLs and download many news pages after only one hop. Web administrators usually publish news pages under only a few paths, such as www.website1.com/news/, www.website2.com/2007news/. So URLs of many news pages in the same location 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. An instant crawler can add a newly discovered URL to its waiting list if it starts with any one of the news URL prefixes.

The rest of this paper is organized as follows: Section 2 introduces earlier research in priority arrangement for 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 updates of existing pages and that of novel page appearance[1][2]. The conclusion is that the web is growing explosively[3] and it is almost impossible to download all novel pages. Web crawlers face a frontier which is

consisted of discovered by 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[4][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][9]. They estimate whether a URL is worth downloading mainly based on its anchor text. Other crawlers[10][11][12][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. Data Set

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	Requested URL	Referrer URL

Navigation history is also recorded. A user accesses the Requested 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 for 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. News Page Properties

Click through data evolution of news pages is different from that of non-news pages. Many news pages are linked by a relatively small number of hub pages. News pages stored together share the same URL prefixes. These properties can be used in discovering novel news pages.

focus.cn is a web site about real estate which publishes both news pages and non-news ones. 3,040 pages from focus.cn are annotated manually, of which 2,337 are labeled news and 703 are labeled non-news. The proposed news page properties are verified with this dataset.

4.1 Click Through Evolution of News Pages

A news page reports recent events. It is attractive only in a short period after it is published. Later, as more and more users are familiar with the event, the page becomes outdated and fewer users are glad to access it. Its daily click through number decreases greatly. In contrast, a non-news page is not related with a recent event and

does not become obsolescent. Its daily click through number does not fluctuate greatly over time. Click through evolutions of the two types of web pages are shown in Fig. 1.

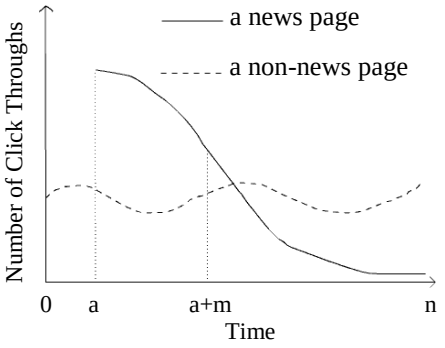

Fig. 1. Click through evolution of a typical news page and a typical non-news page

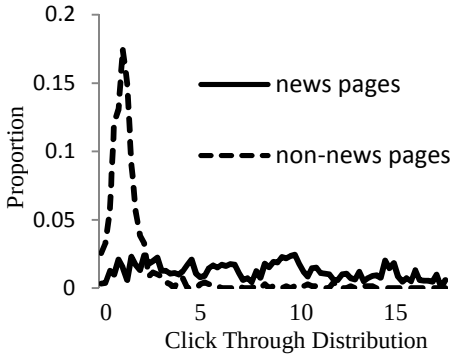

Fig. 2. ClickThroughConcentration distribution

Click throughs of a news page concentrate in the first few days after its birth, while those of a non-news page are more evenly distributed. ClickThroughConcentration, which measures the extent of click through concentration, is defined below:

$$\text{ClickThroughConcentration}(p) = \frac{\sum_{i=a}^{a+m} \text{ClickThrough}_p(i) / \sum_{i=a}^{n} \text{ClickThrough}_p(i)}{m/(n-a)} \qquad (1)$$

where $\text{ClickThrough}_p(i)$ is the number of click throughs of page p on the i*th* day, a is the first day when $\text{ClickThrough}_p(a)$ is non-zero, m is the parameter, n is the width of the observation window, n - a is the length of the observed life of page p, $\sum_{i=a}^{a+m} \text{ClickThrough}_p(i) / \sum_{i=a}^{n} \text{ClickThrough}_p(i)$ is the number of click throughs accumulated in the first m days divided by that accumulated in the observed window. The birthday a of the news page is shown in Fig. 1. The numerator reflects the proportion of click throughs received in the first m days. The denominator is used for normalization. If daily click throughs of a page are evenly distributed, ClickThroughConcentration would be 1. Otherwise, if click throughs concentrate in the first m days, ClickThroughConcentration would be greater than 1.

ClickThroughConcentration of all 3,040 pages from focus.cn are calculated to verify the previous assumption and their distribution is shown in Figure 2. ClickThroughConcentration of 91.9% non-news pages is less than 3 and the average is 1.86. The average ClickThroughConcentration of news pages is 9.11. ClickThroughConcentration of news pages is more widely distributed because pages reporting breaking news get through users rapidly and become outdated soon, while those which are not so attractive need more time to get through users and their daily click throughs are not so concentrated.

As is shown in Figure 2, ClickThroughConcentration of news pages is larger than that of non-news pages. This is a useful feature to discriminate news pages from non-news ones.

4.2 News Pages are Linked by News Hub Pages

A news hub page is a web page which has links pointing to many news pages. It is valuable in discovering news pages. Web page link structure is not available in our data set because page content is not stored. But link structure can be reconstructed from navigation history. Page A links to Page B if a user accessed Page B from referrer Page A. It is true that links which have not been clicked are omitted. But from users' perspective, since they are not clicked at all, there is not much difference whether they exist.

home.focus.cn/, house.focus.cn/ and house.focus.cn/excl/msnhot.php are three major news hub pages. They have links pointing to 1,953 different news pages. A large proportion of news pages can be discovered from these news hub pages.

4.3 Many News Pages' URLs Start with the Same Prefixes

Many news pages cluster in the same news folders. A folder in a website is a news folder if the proportion of news pages in it is large enough.

/news/, /msgview/ and /msn/ are the news folders where most news pages from focus.cn are stored. Their distribution is shown in Table 1. As is shown, three news folders contain 99% news pages and only 4 non news ones.

Table 1. Web page distribution in news folders and other folders

	News Pages	Non-News Pages
In the three news folders	2,322	4
In the other folders	15	699

Additionally, a dynamically generated page, such as website.com/a.asp?p=1, is likely to be a news page if most other pages generated from the same program website.com/a.asp with different parameter values are news pages.

A news URL prefix is a string with which many news pages' URLs start. Pages in the same folder or generated from the same program share the same URL prefixes. Instant crawlers can reach many news pages with a small overhead of non-news ones if they only download novel pages whose URL start with news URL prefixes.

These news page properties discussed above are common in many websites. They can be used to increases the discoverability of novel news pages for instant crawlers.

5. 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.

5.1 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.

5.2 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. 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 dynamic 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 ones directly and indirectly under that node. For example, a 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. 3.

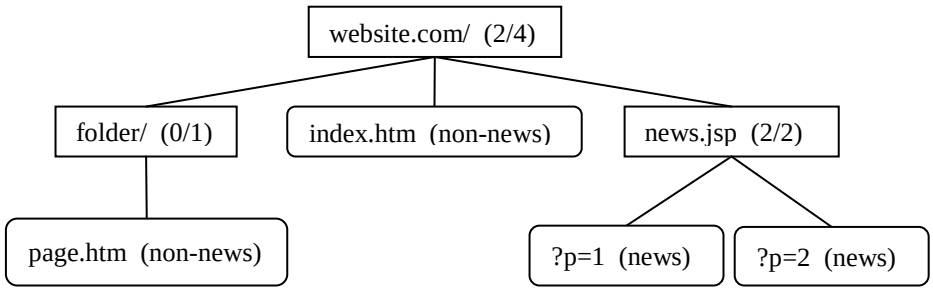

Fig. 3. A URL prefix tree of a sample website

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, then the traverse stops. Otherwise, its children 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.

6. Experiment and Evaluation

An experiment is conducted in the dataset described in Section 3. All pages appeared in the dataset are classified as news pages and non-news ones. A seed URL list is made and news URL prefixes are generated. An instant crawler can start from the seed URLs and decides whether a newly discovered URL is worth downloading based on whether the URL starts with one of the news URL prefixes.

6.1 Experiment

ClickThroughConcentration of all pages are calculated. 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, miss (the number of news pages that are classified as non-news) is 188 and false alarm (the number of news pages that are classified as non-news) is 177. This threshold is applied on all pages and 147,927 are labeled news pages and the 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 threshold, where 0.8 is used. 439 nodes are labeled news nodes. Larger threshold can be used if the bandwidth is limited and that wasted in downloading non-news pages is unaffordable. If an instant crawler has enough bandwidth and wants more news pages, the threshold can be smaller.

6.2 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.

46,210 different news pages are linked by 1,542 homepages of news sites in Sogou’s list, while 79,292 different news pages are linked by 1,542 news hub pages which are included in our seed list. Not all homepages are the best seed URLs. 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/, not its homepage. So finance.sina.com.cn/stock/ 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.

Table 2 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%

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. Instant crawlers can downloads more news pages with less burden of non-news ones.

7. 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. D. Fetterly, M. Manasse, M. Najork, and J. L.Wiener. A large-scale study of the evolution of web pages. Software Practice and Experience, 2004.
2. B. E. Brewington and G. Cybenko. How dynamic is the web? WWW9 / Computer Networks, 2000.
3. Brewington, B. & Cybenko, G. (2000). How Dynamic is the Web, Proceedings of WWW9 –9th International World Wide Web Conference (IW3C2), pp. 264-296.
4. J Cho, H Garcia-Molina. Effective page refresh policies for Web crawlers. ACM Transactions on Database Systems (TODS), 2003
5. SHKAPENYUK, V. AND SUEL, T. 2002. Design and implementation of a high-performance distributed web crawler. In Proceedings of the 18th International Conference on Data Engineering (San Jose,Calif.).
6. Menczer F,Belew R.Adaptive Retrieval Agents:Internalizing Local Context and Scaling up to the Web[J].Machine Learning,2000,39 (23):203-242.

7. Pant G,Menczer F.Topical Crawling for Business Intelligence[C]//Proc 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL).Trondheim,Norway,August 17-22,2003.
8. K. Stamatakis, V. Karkaletsis, G. Paliouras, J. Horlock, et al., Domain-specific Web site identification: the CROSSMARC focused Web crawler, in: Proceedings of the 2nd International Workshop on Web Document Analysis (WDA2003), Edinburgh, UK, 2003.
9. Filippo Menczer , Gautam Pant , Padmini Srinivasan, Topical web crawlers: Evaluating adaptive algorithms, ACM Transactions on Internet Technology (TOIT), v.4 n.4, p.378-419, November 2004
10. J. Cho, H. Garcia-Molina, and L. Page. Effecient crawling through URL ordering. WWW8 / Computer Networks, 30(1-7):161–172, 1998.
11. N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the web frontier. In Proc. 13th WWW, pages 309–318, 2004.
12. Nadav Eiron and Kevin S. McCurley. Locality, hierarchy, and bidirectionality in the web. In Workshop on Algorithmsand Models for the Web Graph, Budapest, May 2003.
13. Serge Abiteboul, Mihai Preda, and Grgory Cobena. Adaptive on-line page importance computation. In Proc. 12th World Wide Web Conference, pages 280–290, 2003.
14.