Board Forum Crawling: A Web Crawling Method for Web Forum

Yan Guo        Kui Li        Kai Zhang        Gang Zhang
Software Division, ICT, CAS
        guoy@ict.ac.cn

Abstract

We present a new method of Board Forum Crawling to crawl Web forum. This method exploits the organized characteristics of the Web forum sites and simulates human behavior of visiting Web Forums. The method starts crawling from the homepage, and then enters each board of the site, and then crawls all the posts of the site directly. Board Forum Crawling can crawl most meaningful information of a Web forum site efficiently and simply. We experimentally evaluated the effectiveness of the method on real Web forum sites by comparing with the traditional breadth-first crawling. We also used this method in a real project, and 12000 Web forum sites have been crawled successfully. These results show the effectiveness of our method.

1. Introduction

Web Crawlers (also called Web Spiders or Robots), are programs used to download documents from Internet. Traditional breadth-first crawling, which is called as TBFC in this paper, is popular used in all kinds of cases. However, for different type of Web sites (e.g. News sites and forum sites), there are so many differences among their organized structures that TBFC cannot crawl all of them efficiently. We believe that different crawling methods should be developed to fit for different type of Web sites. In this work, we deal with crawling Web forum.

Web forum sites have become precious deposits of information, and crawling of Web forum has become more and more important and significant. Generally, for a Web forum site, the target of crawling is to download all posts in the site. Web crawling is a well-studied research problem. Some issues (e.g. crawling Hidden Web [1] and user-centric crawling [2]) have been hot research points. However, to the best of our knowledge, there has been little amount of research on the crawler especially for Web forum.

When a Web forum is crawled by TBFC, Spider Trap and noisy links are main obstacles for precise and efficient crawling. Such troubles are mainly caused by conflicts between the organized characteristics of Web forum sites and the characteristics of TBFC.

The TBFC works as follows: at first it follows all links in the homepage to download all pages linked by homepage, and then follows all links in those downloaded pages to download all pages linked by those pages, until there are no more new pages in the site linked by downloaded pages.

For most of Web forum sites, there are some characteristics as follows, from which we can explain why Spider Trap and noisy links can be encountered:

(1) Most of the Web forum sites are designed as dynamic sites. Most of the information contained in forum sites is usually organized in databases. When two requests which requiring the same piece of content in the database are forwarded to the Web server, the server will return two dynamic Web pages which having the same content but different URLs to the client. Here we call the two dynamic pages as redundancy pages, and the links leading to redundancy pages are called duplicated links. So when the forum sites are crawled by TBFC, there must be a lot of redundancy pages being downloaded and a lot of duplicate links waiting for crawl. As a result, the links needed to crawl become much more than reasonable links of the site, just as if the site had infinite links need to be crawled, which is known as Spider Trap. For the redundancy pages, although they have the same content, but they have different URLs, so the crawler can not eliminate the duplicate links by URL checking and still downloads all of them.

(2) In a Web forum site, there are a lot of noisy links, such as the functional links for users to “print”, and the links of some advertisements. The pages linked by the noisy links almost have no useful information. As a result, crawling the noisy links not only wastes the resource of TBFC, but also lowers the quality of the downloaded pages.

(3) Links in a Web forum site are organized as some levels. To find a post in a board, we have to start from the homepage, and then enter into a board, and then to find the post. Here we should note that in a Web forum site, the most useful information is contained in the deep levels. Since TBFC starts crawling from the
homepage and then continues crawling by extending links, only when the crawler crawls into enough deep levels, the meaningful pages can be downloaded. When a Web forum site is crawled, most popular search engines, which using TBFC, limit the scale of crawling in order to avoid falling into Spider Trap. However, what they have done can usually lead side effect: since they can not crawl enough deeply in the site, the information they gained are certainly only a very small part of the whole information contained in the site.

To solve the above troubles, we provided a method of Board Forum Crawling, which is called as BFC in this paper, especially to crawl Web forum sites. Experimental and practical results showed that, in both theory and practice, it is an efficient and simple method for crawling Web forum.

2. Related work

Crawlers are important component of Web search engines, and as such, their technique details are kept as business secrets. Some issues (e.g. crawling Hidden Web [1] and user-centric crawling [2]) have been hot research points. However, there has been little amount of literature in this area, see [1-5] for a (no exhaustive) list of efforts.

3. Method of BFC

3.1. The Idea of BFC

Most of Web forum sites are constructed by a small set of forum software, such as Discuz!, Dvforum and so on. The forum sites created by the forum software have the similar structure on organization. Our method of BFC exploits the organized characteristics of the Web forum sites and simulates human behavior of visiting posts of the site. Our idea is shown as follows.

To visit a post in a board, human usually starts from the homepage, and then enters into a board, and then to find the post. This process hints that the Web forum is organized structurally. Thus we found that there exist mainly 3 kinds of pages in one Web forum site: homepage and post page and board page.

(1) Homepage is the entrance page of the site; for example, the homepage of site www.ubuntuforums.org is shown in Figure 1.

(2) A post page records some views on a topic provided by authors, and the post page contains exact information of the Web forum site.

(3) A board page contains a link index of some post pages in one board. For example, a board page of site www.ubuntuforums.org is shown in Figure 2. From a board page, we can extract a small link index of part of post pages in one board. And from all board pages of a Web forum site, we can easily create a whole link index of all post pages in the site. By the whole link index, we can download the post pages of the site efficiently and simply. Now the main problem we have to deal with is that: how to get all board pages of one Web forum site?

Figure 1. An example homepage

Figure 2. An example board page

From Figure 1, we can see that the newest board pages of most boards in the Web forum site are linked by the homepage, so that we can extract the links of the newest board pages from the homepage, and we call these newest board pages as board page seeds in this paper. From Figure 2, we can see that one board page always links some other board pages in the same board, and we call those board pages which are only linked by some other board pages in the same board as
subsequent board pages. Thus board pages can be clustered into 2 sets, and one set includes board page seeds, and the other set includes subsequent board pages. From a board page seed, all subsequent board pages in the same board can be recursively gained, and from all board page seeds, we can get all subsequent board pages in the Web forum site.

3.2. Method of BFC

Based on the above idea, we provide the method of BFC to crawl Web forum. The description of BFC is shown as follows (Limiting to the space, the details of the method, such as link clustering based on URL, are not described here.):

Input: a homepage of a Web forum site
Output: most post pages in the site
Method:

Step 1. Extract board page seeds from homepage.
Step 2. For each board page seed, get a link queue of all subsequent board pages in the same board with the input seed.
Step 3. For each queue, download each page in the queue, and identify whether it is exactly a board page and extract links of post pages from the board page. At last create a whole link index of all post pages in all board pages.
Step 4. Download post pages linked by the whole link index gained in Step 3.

An example of part of a whole link index of a Web forum site is shown in Figure 3.

4. Experiments and evaluations

All experiments were performed on a Windows 2000 machine with two 2.4 GHz Pentium IV processors and 4GB memory.

4.1. The quality of the downloaded pages in TBFC

In order to evaluate the quality of the downloaded pages in TBFC, we have used TBFC to crawl a group of sites, and we only crawl no more than 5 level links. The result is as follows:

<table>
<thead>
<tr>
<th>Web forum site</th>
<th>①</th>
<th>②</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.cntong.com/phpbb/">www.cntong.com/phpbb/</a></td>
<td>94193</td>
<td>3452</td>
</tr>
<tr>
<td>csapa.org/phpBB/</td>
<td>155560</td>
<td>7201</td>
</tr>
<tr>
<td>forum.lnnu.edu.cn/</td>
<td>827318</td>
<td>78388</td>
</tr>
<tr>
<td>forum.centrmus.com/</td>
<td>616286</td>
<td>16435</td>
</tr>
</tbody>
</table>

Note:
① Count of downloaded pages
② Count of posts

From Table 1, we can see that the quality of the downloaded pages in TBFC is not very high. Here, posts are counted manually.

4.2. Evaluation of the method of BFC

4.2.1. The Precision and Recall. We provide the Precision of downloaded pages and the Recall of post pages to evaluate the effectiveness of BFC. We suppose that, for crawling a Web forum site, Page_downloaded is the count of all pages downloaded, and Post_downloaded is the count of all post pages downloaded, and Post_all is the count of all post pages in the site. Then the Precision and Recall are computed as follows:

\[
\text{Precision} = \frac{\text{Post}_\text{downloaded}}{\text{Page}_\text{downloaded}} \times 100(1)
\]

\[
\text{Recall} = \frac{\text{Post}_\text{downloaded}}{\text{Post}_\text{all}} \times 100(1)
\]
Recall = \frac{Post_{downloaded}}{Post_{all}} \times 100 \quad (2)

We use Precision and Recall to compare the TBFC and our BFC on some real Web forum sites, the result is shown in Table 2. Here, Post_all and Post_downloaded are counted manually. Here, we only extract post pages from the first 32 board pages in the queue. From Table 2, we can see the Precise of our BFC can reach to 90%, which is much higher than the TBFC. The Recall of our BFC is also higher than the TBFC in most sites.

**Table 2. The Precision and Recall of TBFC and BFC**

<table>
<thead>
<tr>
<th>Web forum site</th>
<th>1</th>
<th>2</th>
<th>BFC</th>
<th></th>
<th>3</th>
<th>4</th>
<th>5</th>
<th></th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>csapa.org</td>
<td>6601</td>
<td>7201</td>
<td>6554</td>
<td>99.3</td>
<td>91.0</td>
<td>2219</td>
<td>33.6</td>
<td>30.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.cntong.com">www.cntong.com</a></td>
<td>2531</td>
<td>3452</td>
<td>2472</td>
<td>97.7</td>
<td>71.6</td>
<td>1352</td>
<td>53.4</td>
<td>39.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.linuxbyte.net">www.linuxbyte.net</a></td>
<td>2148</td>
<td>3020</td>
<td>2097</td>
<td>97.6</td>
<td>69.4</td>
<td>633</td>
<td>29.5</td>
<td>21.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.soking.com">www.soking.com</a></td>
<td>1949</td>
<td>2605</td>
<td>1849</td>
<td>94.9</td>
<td>71.0</td>
<td>357</td>
<td>18.3</td>
<td>11.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>luntan.popo.163.com</td>
<td>50378</td>
<td>142315</td>
<td>49122</td>
<td>97.5</td>
<td>34.5</td>
<td>3597</td>
<td>7.0</td>
<td>2.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>envi.ruc.edu.cn/forum/</td>
<td>1899</td>
<td>2103</td>
<td>1713</td>
<td>90.2</td>
<td>63.4</td>
<td>326</td>
<td>17.2</td>
<td>15.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ①Page_downloaded ②Post_all ③Post_downloaded ④Precision ⑤Recall

4.2.2. The time cost of BFC. We have evaluated the time cost of the method of BFC. The statistics of time cost for creating a whole link index of all post pages of some real Web forum sites is shown in Table 3. Here, we only extract post pages from the first 3 board pages in the queue.

**Table 3. The time cost for creating a whole link index**

<table>
<thead>
<tr>
<th>Web forum Site</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th></th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.discuz.net/">www.discuz.net/</a></td>
<td>1</td>
<td>56</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>forum.dvforum.net/</td>
<td>5</td>
<td>20</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.52magic.net/Boards.asp">www.52magic.net/Boards.asp</a></td>
<td>3</td>
<td>136</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| forum.leoforum.com/cgi-bin/leofo
rum.cgi                          | 23   | 1866  | 756  |      |      |      |
| www.bmforum.com/cgi-bin/npix.p
h                               | 9    | 150   | 30   |      |      |      |
| www.readfree.net/forum/        | 10   | 72    | 17   |      |      |      |
| www.linuxsir.org/forum/        | 120  | 3356  | 500  |      |      |      |
| forum.putclub.com/             | 11   | 520   | 118  |      |      |      |
| www.ofstar.net/                | 1    | 55    | 11   |      |      |      |
| www.molyx.com/forum.php        | 11   | 235   | 40   |      |      |      |

Note:
①Time for Getting Board Page Seeds (s)
②Time for Getting Links of All Board pages (s)
③Time for Getting Link Index of Post pages (s)

From Table 3, we can see that for BFC, the total time cost for creating a whole link index of all post pages is exactly not much.

5. Conclusions

We present a new method of BFC to crawl Web forum. Experiments have shown BFC is an efficient and economical method. It is worth to note that BFC has been used in a real project, and 12000 Web forum sites have been crawled successfully. However our research is mainly based on the Web forum in China where most forums have the similar structure. In the future, we will optimize the method of BFC to make it more efficient and more general for crawling Web forums.

6. References