Intelligent techniques for searching Internet forums

Miloš Pavković[1], Dražen Drašković[1], Lovro Šubelj[2], Slavko Žitnik[2], Dejan Lavbič[2], Marko Janković[2], Jelica Protić[1], Boško Nikolić[1]

[1]University of Belgrade, School of Electrical Engineering
[2] University of Ljubljana, School of Computer and Information Science
E-pošta: {pm115002p, drazen.draskovic}@etf.bg.ac.rs,
{lovro.subelj, slavko.zitnik, dejan.lavbic, marko.jankovic}@fri.uni-lj.si,
{jelica.protic, bosko.nikolic}@etf.bg.ac.rs

Abstract - Web crawlers are one of the most Internet searching techniques today. This paper provides an overview of the most famous forum web crawlers, their performances and disadvantages. Searching problems have been examined predominantly and improvement measures leading to better searching results have been proposed and exemplified by web forums.

Key Words - Web Crawlers, Searching, Forums.

1 Introduction

Nowadays, there are numerous Internet forums dealing with diverse topics. The interpretation of forums' contents in a correct way is of enormous significance for web crawlers. Searching forum contents differs from searching classical web pages [1]. The biggest problem is to match and sort key elements, such as: forums, topics, posts and authors into some logical whole which would be more easily analyzed and sorted later on. An additional problem is also the non-standardized way of representing key elements of the forum. From a technical viewpoint, search algorithms spend large Internet resources of the server on which there are located, as well as the processing time.

A solution advocating the approach of searching which is as modular as possible and is based on intelligent techniques of forum key elements recognition, as well as the optimization in the course of server's resources consumption.

2 Knowledge base and search phases

Knowledge base is necessary for recognizing key elements of the forum. Knowledge base is primarily split into language groups and language areas, and each of these wholes contains sets of regular complex expressions used for recognition of forum technology, forum lists, topics, posts and post's authors.

Today's most popular forum-creating technologies are vBulletin, phpBB, Discuz!, Phorum, YaBB and many others. 95% of the forum located on the web today is exactly generated by means of these already known technologies, while the remaining 5%, which are intended for a specific purpose, may be represented by the sets of the already existing regular expressions.

The forum technology and language are detected first, and then a set of regular expressions pertaining to a concrete language area and production technology [2][3]. Regular expressions within one technology are divided into four groups:

1. Forum's links detection
2. List topics and list dates detection
3. Detecting and singling out all the relevant contents of one post, which can be provided by concrete technology
4. Page detection

The regular expressions in the knowledge base are expanded by special symbols '\digit', '\alpha' and '>>'. The first two represent a number and a letter character, respectively. The third symbol, however, enables the positioning to the concrete characters set which follows it. This characters set may also contain regular expressions within itself. In that way, a powerful tool for positioning within the text has been obtained.

The examples of regular expressions detecting links to forum lists of popular technologies have been given in the Table I.

Table 1. Examples of regular expressions

phpBB	`>>viewforum\.php\?=f\==\digit+`
vBulletin	`>>/forum/vbulletin=\-=!/+/=.+`
Phorum	`.+/list\.php=\?=\digit+>>=`
YaBB	`>>/community/=YaBB\.pl\?board\==.+`
Discuz!	`>>/forum\-=digit+\-=1\.html=`

Searching in this model is broken into three phases:

1. Searching forum lists
2. Searching forum lists' topics
3. Searching posts according to topics

Neither of these phases except for the first one can be initiated until the previous phases have been completed. Forum lists, topics and posts are three different logical entireties having diverse functions. Each of these three logical entireties has an individual algorithm aimed at their searching and it stores the search results in a separate base. Later on, this kind of organization makes possible easier sorting, analysis and search of the obtained data.

2.1 Forum lists search

In the first phase, forum lists are being searched according to knowledge base. The information

memorized by the web crawler include: the forum link, forum title and its ID. The forum link is an element necessary in the next search phase, while the title is necessary for the users to inspect the search results. The forum ID is used as an identifier within the system in which these data are stored. ID parameter is being generated with regard to a forum identifier within the link itself. At the same time, this is a forum list identifier on the forum itself. The record format is XML and is illustrated by means of the following example:

```
<Forums>
 <forum>
  <name>Primer foruma</name>
    <url>
        http://www.primer.com/viewforum.php?f=6
    </url>
  <id>6</id>
 </forum>
</Forums>
```

2.2 Topics search on the forum lists

The second phase entails crawled thread lists and their topics selection. Every thread list has its own activity frequency per topic, and this may vary from one forum to another. The first time the web crawler visits the forum thread list, it collects all the topics from that particular list. A web crawler visits a concrete forum list anew according to the activity frequency in order to detect and gather new topics and posts. This model optimizes the search results by searching only the posts that are more recent than the date of the last visit to that particular forum. In order to achieve this, a web crawler takes into consideration also the dates on the topics list. Topics may be sorted in two ways:

1. according to the last activity (post) date
2. according to the thread creation date

Figure 1 exemplifies, by way of illustration, the forum search algorithm classification division.

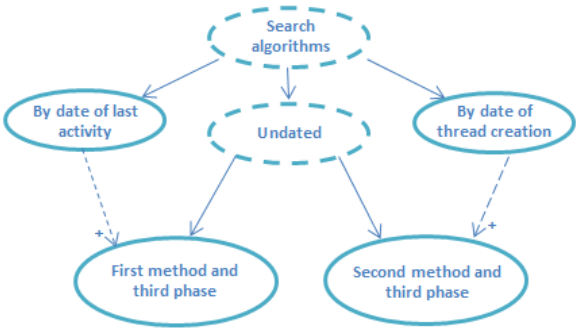

Figure 1. The classification of the analyzed forum search algorithms

Some thread lists need contain neither of these dates regardless of the way in which they have been sorted; and there may be even cases where the list has been sorted according to the last activity, and the dates have been displayed according to creating, and vice versa. An additional problem may lie in the fact that all these thread lists are page-related. This model supports solutions to any of these combinations and makes an effort to optimize it to the maximum extent according to the Internet resources consumption of the server on which it is hosted, and time required to process fetched pages.

The first and simplest method refers to lists, which contains the date and sorted with regard to the last activity. The web crawler will acquire threads, one after the other, the date of which is more recent than the last day visit per the given threads list. The moment it comes across an older date, it ceases searching the whole list, because all the dates that follow are older and all the threads and their posts that follow have already been collected.

The second method is applied for the lists sorted according to the thread creation date. This method supports the fact that the threads older than a certain date *dt* will mostly not be active. The precondition for list visiting cessation is attaining the page which does not contain any topic the creation date of which is more recent than dt. The date *dt* is calculated as the difference between the current time CURRDATE and the parameter MAXDAYS, as exemplified by the equation (1):

$$dt = CURRDATE - MAXDAYS \quad (1)$$

MAXDAYS is the value which is generated after visiting the forum for the first time, and it represents the average duration value of one topic in relation to days. In equation 2, n is the number of topics, while LPD and FPD stand for last activity date and topic creation date, respectively (Last Post Date and First Post Date). Every other visit modifies this parameter depending on the current global activity of a forum and its topics activities.

$$MAXDAYS = \frac{\sum_{i=1}^{n}(LPD_i - FPD_i)}{n} \quad (2)$$

The third method pertains to undated lists. During the first step, the method detects sorting of a list by visiting the topics from a given page and by detecting the date of the posts themselves.

In case of detection of sorting according to last activity date, an algorithm is reduced partially to the first method and, furthermore, it combines the third phase comprising the topic post visit. Topics are visited in a certain order one after the other. Each time a topic is visited in search of the last post date this activity is used in order to collect also the new posts stemming from that particular topic. After the posts have been collected, the algorithm contains the last post date per that particular subject. If a date is older than the last visit date, the other topics visit is being discontinued; however, if this is not the case, than the other topics visit proceeds.

In case of detection of sorting according to thread creation date, an algorithm is reduced partially to the second method in that it combines the third phase of topics post visit, in the same way described for the case of detection of sorting according to last activity date.

The data which are always kept and stored are: a topic link, a topic's title and an adequate date (LPD or FPD). In addition to the link to the posts they contain, the vast majority of forums contain a direct link to the last post within the topics list. If they exist, these links are also kept and stored.

2.3 Searching posts according to topics

The most essential forum parts, i.e. posts, are collected in the third phase. Each post is being processed separately. One post contains the post contents as the main data, or more precisely, it contains the text and connects it with the post creation date and with the author who has created it. If there are additional pieces of information concerning the author, such as the profile picture, forum join date, location, signature or e-mail, these are also brought into connection with the author.
In case of all posts being on one side, the algorithm parses the whole page and discharges that topic. If the posts extend over several pages, then a distinction is made between two cases: when the posts are sorted chronologically according to the list date and threaded lists. Threaded lists are special type lists which may contain a reply to the already existing post. Nevertheless, posts are not sorted chronologically according to the date on these lists.

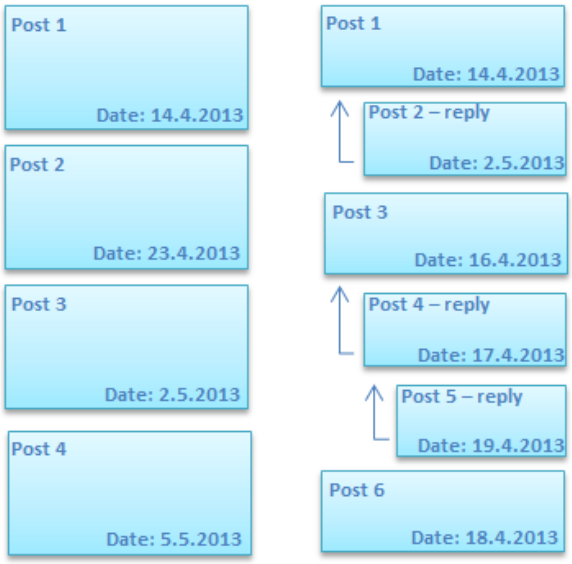

Figure 2. Searching posts within a single topic

When posts are sorted chronologically, a distinction is made between two methods that have been optimized so as to fetch as smallest number of pages as possible thereby reducing the consumption of Internet resources and processor's time.
1. If the link to the most recent post has been successfully found in the second phase, this link is being accessed first. The posts are then collected from this link sequentially one after the other until the post creating date older than the last date forum visit to this thread has been reached. Then, the crawling process is terminated. This method is also called BCKW or Backward method.
2. If the link to the most recent post has not been found in the second phase, then the first page of thread is being accessed. All crawlers collects the first page visited posts and more recent ones, if they exist. Subsequently, the link to the last page topic and if the next approach has been found then this link is accessed. Furthermore, all the pages are visited backward and all the posts are collected one after the other, until the post creating date older than the last forum visit date has been reached. At that time, the browsing is terminated. If the search backward reaches the first page, then this page is not browsed. This method is also called FLP or First->Last->Previous method.

In case of threaded list or impossibility of applying the FLP/BCKW methods it is necessary to visit all the pages for there is no accurate way of determining when visiting should stop. The diagram of the method is given in Figure 3.

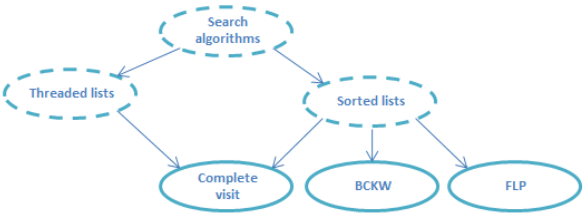

Figure 3. Search algorithms methods for sorted lists

3 Performances analysis

The performances analysis was based on the specimen comprising 100 web forums pertaining to different geographic locations, diverse languages and creation technologies. The test, which lasted six months, was carried out on two models, these being: forum model and standard search model. The searches were released once every 10 days and then the consumed Internet traffic and the total time needed for all 100 web forums were searched.

If looked at the table, a conclusion may be reached that this method is almost identical only during the first access to some forum, because both first and second model must search the complete forums. Economizing on performances and traffic is only seen after the next approach whereby the forum model searches for the most recent forum contents only, while the standard model searches also the older contents, which have been already visited during one of the previous accesses.

Table 2. Results of the search analysis

PERIOD	Forum model		Standard model	
	Search duration (measured in hours)	Consumed traffic	Crawl time (measured in hours)	Consumed traffic
The first access	182	13.1 GB	176	13.8 GB
After 10 days	12	1.7 GB	140	10.2 GB
After 20 days	12	1.65 GB	137	10.8 GB
After 30 days	15	2.2 GB	130	10.3 GB
After 40 days	14	2.1 GB	131	10.3 GB
...	...	...	...	...
After 180 days	18	2.5 GB	80	7.2 GB

4 Conclusion

Nowadays there is a vast number of information on the web. It is not only important to search as large number of Internet sites as possible, but rather to pay attention to the way these sites are being searched. Forums, as one of the greatest human information resources, have become the focus of interest of large companies. On the one hand, the model represented in this paper seeks to represent the searched data in the best possible form, while, on the other hand, it tries to spend the minimum of Internet resources and thus economize on processor's time. The data collected and sorted in this way are more convenient to analyze than the data obtained by the standard search method, while being unloaded of the unnecessary elements from forum pages that are insignificant.

5 Achnowledgements

The paper has been supported by Ministy of Education, Science and Technological Development, Republic of Serbia, and Ministry of Education, Science and Sport, Republic of Slovenia, as bilateral cooperation project »Intelligent information searching based on ontology« project number 651-03-1251/2012-09/38, 2012-2013.

Bibliography

[1] Y. Yang, Y. Du, Y. Hai, Z. Gao, "A Topic-Specific Web Crawler with Web Page Hierarchy Based on HTML Dom-Tree," Asia-Pacific Conference on Information Processing, 2009, pp. 420-424, DOI: 10.1109/APCIP.2009.110

[2] J. Jiang, X. Song, N. Yu, C. Lin, "FoCUS: Learning to Crawl Web Forums," IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 6, June 2013, pp. 1293-1304

[3] A. Sachan, W. Lim, V. Thing, "A Generalized Links and Text Properties Based Forum Crawler," IEEE International Conferences on Web Intelligence and Intelligent Agent Technology 2012, pp. 113-120, DOI: 10.1109/WI-IAT.2012.213