Mathematical and Software Engineering, Vol. 1, No. 1 (2015), 25-30
Varɛpsilon Ltd, http://varepsilon.com

System for Information Extraction from News Sites

Tihomir Stefanov

Department of Information Technologies
University of Veliko Tarnovo
Veliko Tarnovo, Bulgaria

The present paper deals with a system for crawling and content extraction from news sites. The system of web crawlers extracts textual and graphic information and checks for multimedia content availability. A part of the programming code and the database have been presented.

Subject Codes (ACM): H.2.6, H.2.8, D.2.m.

Keywords: source code, information extraction, databases, information sites.

1 Introduction

What options should a news website offer to have priority over hundreds of informational websites, fighting for the attention of readers and advertisers? In response to this issue, the present system of evaluation and analysis of news sites has been created. The developed system is web-based and fully automated. It has been created for the purpose of crawling, extraction, storage and processing of textual and graphic information from news sites. The system evaluates four news sites in terms of 22 criteria, preset by the media experts searching an answer to the question: "What makes a website successful?" The evaluation methodology is focused on two basic criteria: effectiveness and client satisfaction, taking the user qualification profile into account [1], [2], and [3].

In the model introduced in the present paper, more than 5,000 news articles from four specific sites have been extracted and compared.

2 Methods and Algorithms

The system index file /index.php displays the four sites which the system crawls and evaluates. Each site has id (identification number), used in the processing of the internal links of each of the news sites at a later stage.

The /site-details.php script presents details of the processed internal url-addresses of each news site. Id of the website shall be passed as GET variable (the same id as displayed in /index.php). Beneath them are: table of the processed url-s, title, contents and words count of the news text. They do not provide specific information, but rather exercise control in how the system works, if it needs any adjustments or adding of exceptions for crawling specific sites.

id	siteid	url
30481	4	http://chernomorskifar.com/date/2014/03/12
30482	4	http://chernomorskifar.com/date/2014/03/13
30486	4	http://chernomorskifar.com/date/2014/03/03/page/2
30484	4	http://chernomorskifar.com/date/2014/03/04
30485	4	http://chernomorskifar.com/date/2014/03/03
30488	4	http://chernomorskifar.com/date/2014/01/13/page/2
30487	4	http://chernomorskifar.com/date/2014/01/13
30489	4	http://chernomorskifar.com/date/2014/01/09
30492	1	http://www.dnesbg.com/krimi/spipaha-piyan-tiradzhi...
30502	1	http://www.dnesbg.com/obshtestvo/studenti-ot-diplo...
30507	1	http://www.dnesbg.com/goreshti-novini/srednovekovn...

Figure 1. Table 'temp'

After entering a new site into the system, the id column is automatically generated and becomes an id number. The 'name' and 'url' are entered manually by the operator, the date and time columns are control information, which the system changes upon entering and editing the sites. The 'outlinks' is filled in by means of a function set in functions. php. Its function is to check how many external links point to the entry in the ' url ' column of a particular row in Table ' site ', this information being receivable at the following address: http://ajax.googleapis.com/ajax/services/search/web?v=1.0&filter=0&q=url: {url from the database}

```
CREATE TABLE IF NOT EXISTS `temp` (
  `id` int(255) NOT NULL AUTO_INCREMENT,
  `siteid` int(1) DEFAULT NULL,
  `url` varchar(255) NOT NULL,
  PRIMARY KEY (`id`),
) ENGINE=MyISAM  DEFAULT CHARSET=utf8 AUTO_INCREMENT=33255 ;
```

After entering a new website for the first time, the home page is added as an entry both in the Table ‘site’ and in the Table ‘temp’ (Figure 1). The role of this table is to store url-addresses to be crawled by the system. The internal links of some sites are relative (/index.php) rather than absolute (http://novjivot.info/index.php), and before being saved into the database, they are processed by the system.

After that, the system takes the url with the lowest id in this table, processes the webpage content (separating the news title, the news contents, the article photo ...) and checks if the internal links are not already saved in the Table 'temp'. If they are unknown to the system, they are added as new entries at the end of the table for further processing.

In Table 'pages' stores information about the processed url addresses (fig. 2). In the column 'id' is given the id number of the table entry. The 'siteid' is the identifier showing which website the specific url belongs to. 'title' contains the title of the news text (not that of the HTML document) and 'category' – the category to which the news text refers according to the structure of that particular site.

'items_count' is a column containing the news texts count that refer to that specific url.

In the 'content' the whole news content is being stored, free from HTML formatting - paragraph tags, consecutively/non-consecutively numbered list, header (1-6) and new line are removed, so that they do not hinder the counting of words in the text. The column ‘url’ shows the url of the specific news item. The column ‘date’ shows the date and time of news item publishing, the column 'create_date' – date and time of the online news crawling by the system, in the column 'update_date' – a re-crawl date and time of the webpage, if any is necessary. In the column 'words_count' is stored the number of words in the news, and in the 'newsimage' is the URL of the photo illustrating the contents, if any.

+ Options

id	siteid	title	category	items_count	content	url	date	external_links	create_date	update_date	words_count	newsimage
77	2	27-годишна жена се самоуби скачайки от 12-я етаж и...	sinya-lampa	1	27-годишна жена слоши край на живота си, скачайки ...	http://www.struma.com/sinya-lampa/27-godishna-jena...	2014-03-18 15:16:00	0	2014-03-19 12:23:51	NULL	105	http://www.struma.com/_thumb.php?tf=uploads/image/...
78	2	3,4 г. затвор за бизнесмен, сгазил с джипа си рабо...	sinya-lampa	1	37-годишният бизнесмен К.К., който преди близо 8 м...	http://www.struma.com/sinya-lampa/3-4-g-zatvor-za-...	2014-03-18 15:12:00	0	2014-03-19 12:23:53	NULL	130	http://www.struma.com/_thumb.php?tf=uploads/image/...
79	2	Пуснаха Брендо под домашен арест	sinya-lampa	1	Софийския апелативен съд пусна от ареста сочения з...	http://www.struma.com/sinya-lampa/pusnaha-brendo-p...	2014-03-18 15:05:00	0	2014-03-19 12:23:56	NULL	200	http://www.struma.com/_thumb.php?tf=uploads/image/...
80	2	Скандали в елитно училище в Благоевград. Колектив...	obshtestvo	1	Спешна жалба, подписана от колектива на професиона...	http://www.struma.com/obshtestvo/skandali-v-elitno...	2014-03-18 12:26:00	0	2014-03-19 12:24:00	NULL	132	http://www.struma.com/_thumb.php?tf=uploads/image/...
81	2	Десятте претендентки за "Жена на годината" - Благо...	hailaif	1	Станаха известни имената на десетте номинирани бла...	http://www.struma.com/hailaif/desette-pretendentki...	2014-03-18 18:16:00	0	2014-03-19 12:24:03	NULL	216	http://www.struma.com/_thumb.php?tf=uploads/image/...
82	2	БСП Благоевград изказа възмущение от министър Кр. Вит...	politika	1	Министърът на външните работи Кристян Вигенин прис...	http://www.struma.com/politika/bsp-blagoevgrad-nah...	2014-03-18 17:26:00	0	2014-03-19 12:24:06	NULL	130	http://www.struma.com/_thumb.php?tf=uploads/image/...
83	2	Безценно съровище пътува към Благоевград. Специал...	obshtestvo	1	Специалисти от историческия музей в Благоевград ку...	http://www.struma.com/obshtestvo/bezcenno-sukrovis...	2014-03-18 17:55:00	0	2014-03-19 12:24:09	NULL	119	http://www.struma.com/_thumb.php?tf=uploads/image/...
84	2	Арестуваха трима "бизнесмени" в Сандански, изкупув...	krimi	1	Полицията в Сандански задържа преди час трима &quo...	http://www.struma.com/krimi/arestuvaha-trima-bizne...	2014-03-18 18:21:00	0	2014-03-19 12:24:11	NULL	60	http://www.struma.com/_thumb.php?tf=uploads/image/...
85	2	Важно: Не пийте кафе между 8 и 9 часа сутринта! ...	zdrave	1	Ако сте от тези, които си събуждат по-рано сутрин ...	http://www.struma.com/zdrave/vajno-ne-piite-kafe-m...	2014-03-18 10:11:00	0	2014-03-19 12:24:14	NULL	345	http://www.struma.com/_thumb.php?tf=uploads/image/...
86	2	Община Дупница търси финансиране за ремонт на стад...	obshtestvo	1	Община Дупница и футболен клуб “Марек”...	http://www.struma.com/obshtestvo/obshtina-dupnica-...	2014-03-18 15:25:00	0	2014-03-19 12:24:18	NULL	190	http://www.struma.com/_thumb.php?tf=uploads/image/...
87	2	Опасна комбина "снаха-свекър" в	obshtestvo	1	Баща пусна жалба за закана за убийство	http://www.struma.com/obshtestvo/opasna-	2014-03-17 13:41:00	0	2014-03-19 12:24:21	NULL	206	http://www.struma.com/_thumb.php?tf=uploads

Figure 2. Table 'pages'

The database is automatically filled in by the system. The work is carried out by several scripts where the tasks are generally laid down, and some further scripts setting specific selectors, that differ in each of the investigated sites.

/Controller-XXX.php script establishes a connection with the Table 'temp' and takes the url with the lowest id. After checking which site the link belongs to, controller.php calls out a file to crawl it, a file different for each site. This is necessary on account of the difference in the selector checks performed in each site, of the title selectors, of the text selectors and of the news photo selectors. Besides, some sites are further equipped with CMS systems – blogs, advertisement management systems, etc.

After processing the URL with the relevant controller.php script, controller.php reloads automatically and performs the same tasks with the next url in the Table ' temp '.

/single-page-*.php - as an illustration of the work of scripts processing URL-s from the Table 'temp', single-page-dnesbg-XXX.php is used.

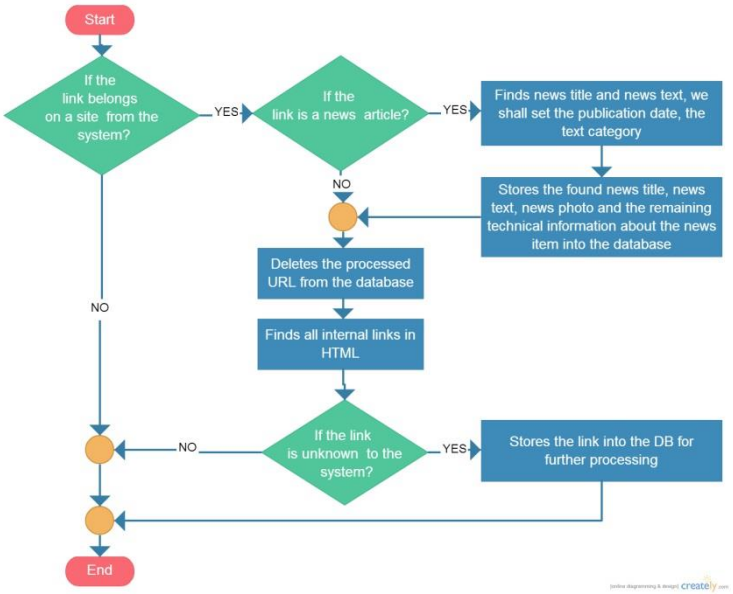

Figure3. Work of single-page-dnesbg-XXX.php script

The work of the script's single-page-dnesbg-XXX. php starts with downloading the url-address with siteid = 1 from the database, which is with the lowest id. (controller. php has already checked if it is suitable for this script, i.e. whether it is part of the dnesbg.com site). This is followed by the creation of a new object, where the URL contents are loaded. This object ($html) contains the entire HTML code of the specific news webpage. The object is defined in the free library parser/simple_html_dom. php. There, are also described the php functions allowing to seek the necessary HTML elements.

```
$html = new simple_html_dom();
$html->load_file($url);
if ((false !== strpos($url, 'http://www.dnesbg.com/'))
&& (false === strpos($url, '?s='))
  && (false !== strpos($url, '.html'))
&& (false === strpos($url, 'advertisements'))
&& (false === strpos($url, 'facebook.com')))
{
```

If the URL matches the preset parameters, it checks the html for the appropriate classes and id's, having text, title, picture:

```
$page_title = $html->find(".entry-title", 0)->plaintext ;
$sadarjanie = preg_replace('/ Споделu $/', '', $html->find(".content-narrow p", 0)->plaintext);
$news_contents = preg_replace( '/\s+/', ' ', $sadarjanie ) ;
$words_count = word_count($news_contents);
```

The word count is performed by means of the word_count function entered in /functions.php

If a content-narrow class containing a photo is available, a link to it shall be retained in $news_image variable. Otherwise □, NULL value shall be set:

```
if ( is_object($html->find(".content-narrow img", 1))) {
$news_image = $html->find(".content-narrow img", 1)->getAttribute('src'); } else {
$news_image = NULL;
}
if (preg_match("/http:\/\/www.dnesbg.com\/([^\/]+)\/.*/", $url, $matches)){
  $category = $matches[1];
} else {
  $category = "index";
}
```

It takes the article publishing date from the .entry-date class and turns it into a table storage friendly form:

```
$date = $html->find(".entry-date", 0)->plaintext;
$date = rtrim(parseBgDate($date), "-");
$time = $html->find(".entry-meta a", 0)->getAttribute('title');
$date = $date.' '.$time.':00';
$date = date('Y-m-j G:i:s',strtotime($date));
```

Sets data and time value of the $sega variable at the moment, in a table storage friendly format:

```
$sega = date("Y-m-d H:i:s", time());
```

Adds the obtained results to the Table ‘pages’ thus completing the block code implementation, which has been fulfilled if the url meets the criteria of a news article.

```
}      // Followed by url deletion from Table ‘temp’:
  $sql2 = "DELETE FROM `temp` WHERE url= '$url'";
// For each link of the webpage HTML the following is fulfilled:
  foreach($html->find('a') as $link)
{
                $vrazka = $link->href;
```

Performs a check, if the link is internal, if it makes a connection to news webpage, etc:

```
                if ((false !== strpos($vrazka, 'http://www.dnesbg.com/'))  && (false === strpos($vrazka, '?s=')) && (false === strpos($vrazka, '.jpg')) && (false === strpos($vrazka, 'advertisements')) && (false === strpos($vrazka, 'facebook.com'))) {
        // Performs a check if the is not already available in the Table ‘pages’
                $sql = mysql_query("SELECT * FROM `pages` WHERE `url` = '$vrazka'");
                $rows = mysql_num_rows($sql);
                                if($rows == 0){
                  // Checks if it is not available in the Table ‘temp’ as well
                    $temporary = mysql_query("SELECT * FROM `temp` WHERE `url` = '$vrazka'");
                        $rows_temp = mysql_num_rows($temporary);
                            if($rows_temp == 0){
                          $add_to_temp = mysql_query ("INSERT INTO `temp` (url, siteid) VALUES( '$vrazka', '1' )");
                            if(! $add_to_temp ){die('Could not enter data: ' . mysql_error()); }
                              }
                        }
          }
        }
```

In Figure 3, block model of the presented algorithm is illustrated.

3 Results and Discussion

The system thus established provides the following statistics for each of the studied sites: most frequently used words, time and date of the publication of each news item, number of words in the news article, number of articles per day, number of photos in the publication. The collected statistical information is displayed in graphic form. The database contains other entries, allowing for further research.

4 Conclusion

The introduced system is made up of searching scripts and a relational database [4]. It allows an objective evaluation of the news sites state in terms of their actuality, information diversity, overall coverage of the presented event, user friendliness of contents layout, easier access to information search, user attendance and many other features in comparison with existing competitive publications. Additional statistical information is displayed in graphic form, which is useful for owners of electronic media.

References

[1] Toleva–Stoimenova, St., Christozov, D., Informing via Websites: Comparative Assessment of University Websites, Issues in Informing Science and Information Technology (IISIT), Vol. 10, 525-537, 2013.

[2] Stefanov, T., Methods for Assessing Information Sites, XLVII International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST'12, Bulgaria, Veliko Tarnovo, 28 – 30 June 2012, Vol.2, p. 455 – 458.

[3] Stefanov, T., Tsvetkov, D., A Model for Evaluation of Regional Electronic Media in terms of Efficiency Criteria and User Satisfaction, Collection of Writings 'Days of Science 2014', Veliko Turnovo, 2014, in press.

[4] Georgieva-Trifonova, Ts., Stefanov, T., Applying linked data technologies for online newspapers, International Journal of Advanced Computer Science and Applications, Vol. 6, No. 5, 2015, p. 29 – 33, Digital Object Identifier (DOI) : 10.14569/IJACSA.2015.060505, The Science and Information (SAI) Organization, ISSN 2156-5570.