DOI 10.34658/jacs.2017.1.23-37

JOURNAL OF APPLIED
COMPUTER SCIENCE
Vol. 25 No. 1 (2017), pp. 23-37

Generalized Structure of the Algorithm for Automated Detection of Non Relevant and Wrong Information on Web Resources

Mykola Dyvak[1], Andrii Kovbasistyi[1], Petro Stakhiv[2], Piotr Lipiński[2]

[1]*Ternopil National Economic University*
Department of Computer Science
Chekhova str. 8, Ternopil, Ukraine, 46003
mdy@tneu.edu.ua, kov_and@ukr.net

[2]*Lodz University of Technology*
Institute of Information Technology
Wolczanska 215, 90-924 Lodz
petro.stakhiv@p.lodz.pl, piotr.lipinski@p.lodz.pl

Abstract. *In this article the algorithm for automated detection of non-relevant or wrong information on websites is introduced. The algorithm extracts the semantic information from the webpage using third party software and compares the semantic information with the reliable resources. Reliable information is identified by the means of majority voting or extracted from reliable databases.*
Keywords: *Semantic analysis of content, parsing, the architecture of software systems.*

1. Statement of the problem

The development of Internet technologies generated large amount of websites. Unfortunately, websites often contain non-relevant or wrong information, which

leads to a sharp drop in web sites traffic [1]-[5]. What is more, these defects are replicated on other websites, business cards, etc., which are used by many organizations and institutions. To correct these errors manually a large number of staff must be involved. Therefore, automatic content analysis tools are required, which could identify non-relevant or wrong information on the websites of organizations.

In [6] content analysis using automatic detection of non-relevant or wrong information on the site card is examined. In this approach information from website of the organization is converted into the format suitable for comparison with other databases available in the organization. This approach was tested by the authors for automated search of non-relevant or wrong information on a number of websites of higher education institutions. Unfortunately, this approach cannot be used for a broad group of websites because it is not always possible to access internal databases of organizations. On the other hand, setting a global task of analysis and content detection of wrong or non-relevant information is quite complicated. This can be achieved by creating intelligent systems based on ontological approach. To accomplish this task it is necessary not only to establish certain characteristics of existing and future algorithms for analysis of content, but also to identify non-relevant or wrong information.

2. The analysis of typical content sites cards

In order to develop tools for content analysis the following parameters should be taken into consideration:

1. The structure of web pages;
2. Features of mark-up language;
3. The language of representation of the content.

Site structure is a logical mark-up and physical connection of pages of the site and layout of the design elements due to standards of development sites [7]. Let us consider the types of structures of the sites that are most typical for business cards [8]. The linear structure of the site: it is the most basic structure of the site. Web-pages are placed one by one, the user must access them in the following order presented below: In the linear structure there is no separation of content (pages) by levels. All pages on such sites are on the same level and are presented in the same way to every visitor. It is easy to realize that html pages are organized in

Figure 1: Linear structure of the site

this way. Each webpage may contain references to other pages. The linear site structure has the following features: focus on one product, the ability to put a large amount of specialized or marketing information, obvious process of selecting a product basing on its specifics. The usage of websites with linear structure is clearly limited. It can be used to image sites (business cards) and online tutorials.

The linear structure with alternatives: this kind of structure is very similar to linear, with the only difference that users have more options to find information - or rather, a choice between two branches. For example, when it is divided for corporate and private clients. The mentioned structure is mostly used to fill in the

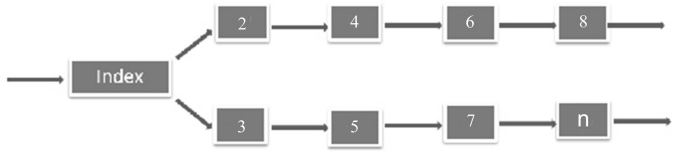

Figure 2: Linear structure of the site with alternative variant A

registration forms. In this case, all visitors start the registration form from the same starting page and continue filling next pages basing on the information provided on previous pages.

Next option are linear structures with an alternative, variant B. In this structure the visitors can visit same page despite making other selections on previous pages. The linear structure of the site with the branches: this structure allows the switching from one page to another in a specific sequence, as in a linear structure. However, visitor can always switch to a different branch, and then come back.

The main advantage of the linear structure with branches is relatively simple algorithm for changing it to linear structure. It is sometimes necessary to improve navigation when the webpage grows. This structure is clearly more complicated.

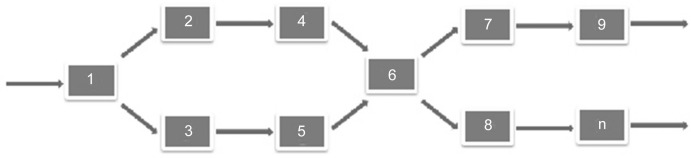

Figure 3: Linear structure of the site with an alternative variant B

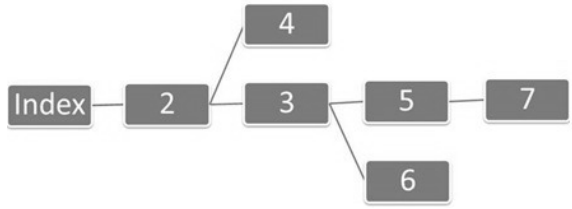

Figure 4: Linear structure of the site with branches

It is peculiar to small corporate resources, websites, business cards, some author's blogs. Typically, there are also no partitions and there are some static pages and links available on a homepage. This navigation system is simple and intuitive. All pages are available in only one or two "clicks". A typical example of a site with a similar structure is a business card of a company.

Lattice structure of the site is one of the most complex structures, where all the documents are located in different branches. However, the visitor can easily navigate between them horizontally (left or right between the branches at different levels) and vertically (top to bottom).

This type of structure is typical for most article directories or links. Nevertheless, lattice structure can easily confuse not only the user when searching for information, but also the webmaster during placing the content. It is also difficult to create and customize as using this structure requires taking avantage of a large number of hyperlinks. Therefore lattice structures are used for large sites.

Next structure which is used for webpage creation is a tree structure. When the webpage has such a structure the user can access any section, subsection and

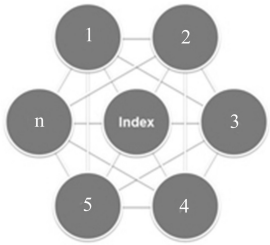

Figure 5: Grid structure of the site

specific page (document) from a home page or from any other page. This site structure is often used by many web masters as the most appropriate. The main

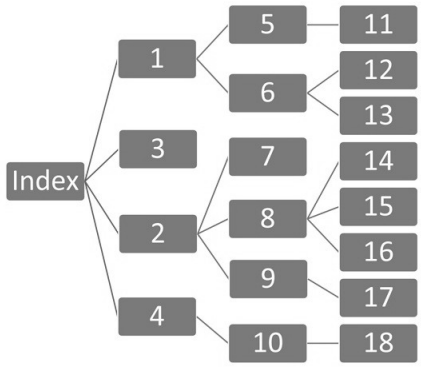

Figure 6: Tree structure of the site

advantage of tree structure of the site is versatility. This structure is suitable for any type of site (home web page, website of organization, corporate site, portals or directories). Such benefits as easy navigation and greater flexibility should also be mentioned. Such structures are difficult to create using HTLM and more advanced tools are required. The disadvantage of tree structure is that it is very difficult to keep balance between "deep and wide". If the "tree" of the site will grow, visitors will have to go through many pages. If the tree is very wide, visitors will have too many choices and it will be difficult select the desired branch. Thus, the webpage

which has tree structure should be constantly monitored in terms of depth and width.

Yet another is a mixed site structure which includes two or more abovementioned structures. Because of the complexity it is rarely used. This structure is more complex than all considered before. All pages are placed in different branches. But the user has the opportunity to navigate between them not only vertically (up and down), but also horizontally (between the branches at different levels). It is mostly used only in catalogs.

The navigation between the branches on deeper levels is established by references to categories in other sections. The automatic content analysis algorithm

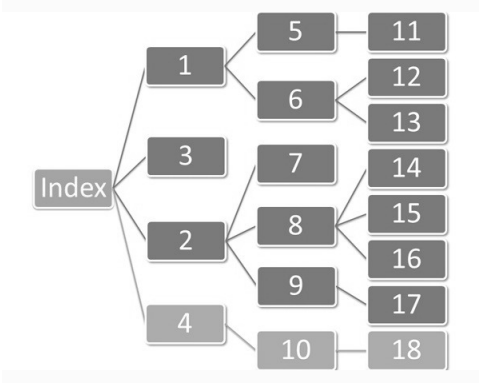

Figure 7: Mixed site structure

developed here uses the abovementioned site structures in the proces of site content analysis, see [9]-[11].

3. The HTML parameters which are taken into consideration

The automatic content management system requires semantic structure of the site to be analyzed. Algorithm for performing semantic analysis is based on the structure of the web page. The majority of web are pages created using HTML (Eng. HyperText Markup Language - hypertext markup language documents), or (XHTML). HTML document is processed by browser and displayed on the screen.

HTML language uses a number of HTML tags to create the structure of the

webpage. Each HTML document contains the following elements: start tag: <html> and end tag: </html>.

Inside the document there are two main blocks: <head> and <body>:

```
<html>
  <head> </head>
  <body> </body>
</html>
```

In the first part of the document the block <head> contains elements that describe the document. They do not appear in the browser.

In Block <head> may contain:

- header document <title>, which sets the title of the page in the browser;
- meta tags - title, description and document keywords <meta>;
- stylistic link to the document file <link>;
- links to JavaScript files used in the document (the tag <script>);
- description of the author of the document.

The second section <body> contains elements of the web page:

```
<html>
<head>
   <title> My page </title>
</head>
<body>
   <h1> Page title </h1>
      <p> Page description </p>
</body>
</html>
```

HTML tag – is an element of HTML page layout. Tags are opened by marking tag with "<" and ">" and closed with same tag marked with "</" and ">"). Some of them do not have close tag - . They also can have the form of a block or string. Block tags begin with a new line and are transferred to the next item on a new line - <p>, . String items are displayed in the same line - , <b>.

The <body> block contains content that is displayed by the browser. Block <body> may contain the following HTML tags: <Table> - table; <a> - reference; <Img> - the picture; <Div> - block element without execution; <P> - a paragraph indent; <Span> - string element without execution; <Form> - form; <Ul>, <ol> - list; <Input>, <textarea>, <select> - form elements; <H1> - <h6> - headers; <B>, <i>, <u>, - string design elements for text - bold, italic; <Audio>, <video>, <canvas> - multimedia elements.

The content of the page (article, information) also requires structuring and styling. To do this, the text is divided into paragraphs <p>, the tables <table>, lists <ol>, <ul>, headings and subheadings <h1>, <h2>, insert links and pictures <a> <img> are used. To highlight important text it is using bold (tag <b>,), italic text (<i>,) or emphasis (<u>). To highlight blocks of text letter spacing or icons can be used.

Consequently, all sites have the same HTML markup that enables to show HTML document in a browser. The content is placed in the block <body>, but also these block tags can contain different names and sequences. Blok <body> also may contain different names which have some meaning that substantially complicates the analysis of content [12].

4. Characteristics of existing systems for analysis of the website content

The structures of webpages, features of the markup language and features of building the markup languages, make it possible to formulate the basic measures for automatic content analysis which allow to analyze typical sites structures. To date, there are a number of systems for analyzing the website content that can serve as a relevant prototypes.

Content analysis is a method of qualitative and quantitative comparison of the content of sites in order to identify or adjust various differences reflected in these sites. The following objects are taken into consideration: the word, proposal, theme, idea, author, character, text.

To create the semantic representation of the webpage DataCol parsing tool is used. This program is available for free in the demo version. With DataCol, it is possible to set up a large number of parsing parameters, including: parser of search engines; parser content; parser Google; parser e-mail; parser online stores; Yandex Store parser; parser ads; parser SEO-parameters; parser music, pictures and other

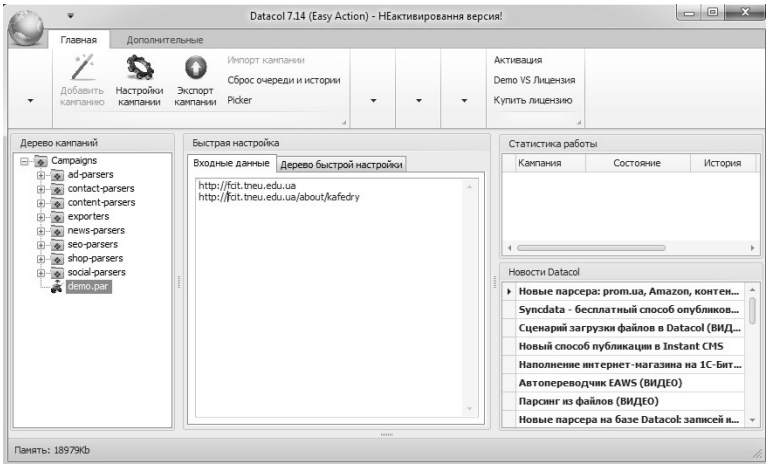

Figure 8: The main window of the program system DataCol

files; parser forums; parser proxy addresses; parser external and (or) internal links from the site.

The resultant information after parsing can be exported to various formats: CSV, Excel, TXT, MySQL, DLE, WordPress, Joomla and others. The results of content analysis can be analyzed using Content Downloader system. This system

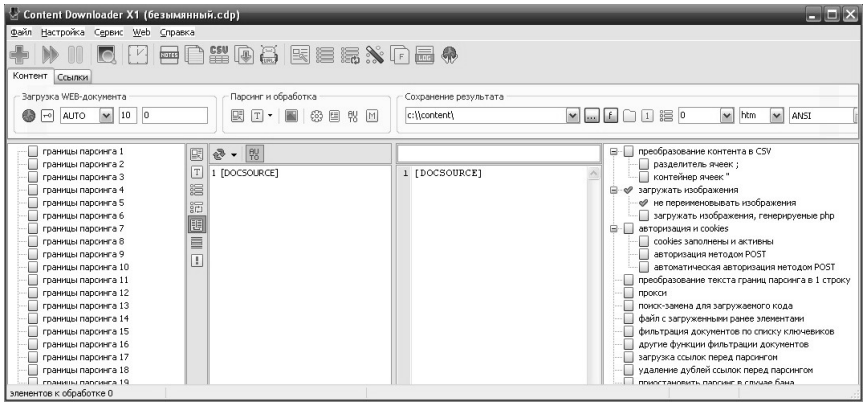

Figure 9: The main window of Content Downloader software system

allows:

- parsing the products of online store in the CSV table (with any desired set of columns output); parsing
- articles with pictures and files (such as files torrents, flash games or abstracts) in htm or txt formats;
- parsing hidden information available only after the click, for example, on the "show the number" or "Show contacts" (requires ULTIMATE type of license, which includes the addition WBApp, which will simulate clicks and buttons);
- parsing hidden information available after the authorization;
- parsing any reference of the site (1 million) that match specific filters;
- parsing any parts of the code WEB-documents and their conclusion in the right format for you;
- parsing XML-Sitemaps.

Options for storage: in one file / multiple files; extension: CSV (with any given column), htm, txt, php, MySQL. The features of content analysis system Sjs-parser. Parser Features: Full grabbing site; partial grabbing; parsing on labels; URL parsing for template; work with filter; grabbing article formatting and images; text cleaning of unwanted characters; removing unnecessary meta tags; parsing configuration file; setting its depth; saving results in formats TXT, CSV, WPT, Zebrum Lightweight and others.

All considered systems have the ability to save results in different formats and set limits of parsing. In addition, they are commercial and have closed source code, which does not allow to use it directly in the system for correcting non-relevant or wrong information in websites.

5. Architecture of the system for content analysis

Based on the results of the analysis performed using existing systems of content analysis of websites we can formulate the basic requirements for automated algorithms for detection of non-relevant or wrong information. Parsing systems should: allow to set URL filter, be able not to parse extra pages, have parsing depth parameter, allow to download the content, be able to perform parallel processing

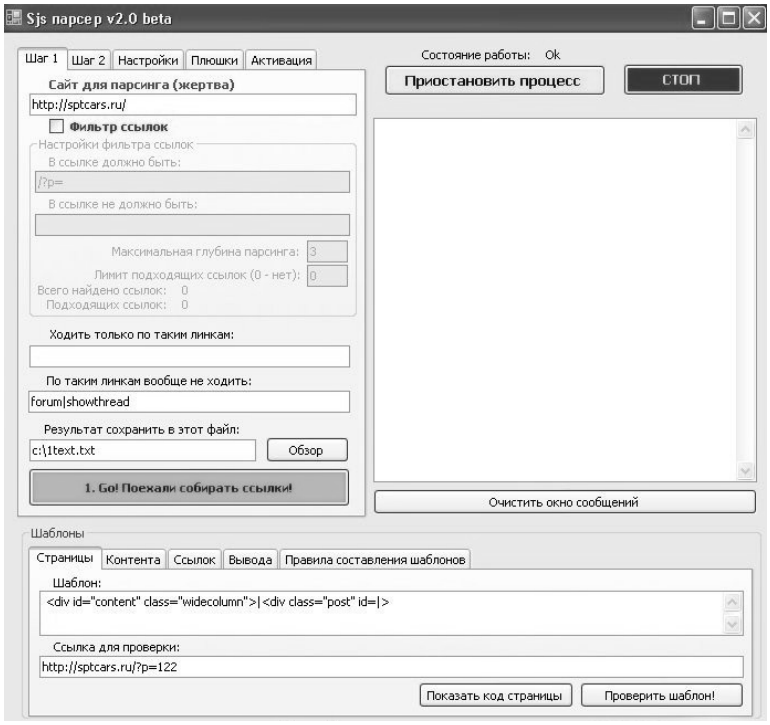

Figure 10: The main window of the Sjs-parser program system

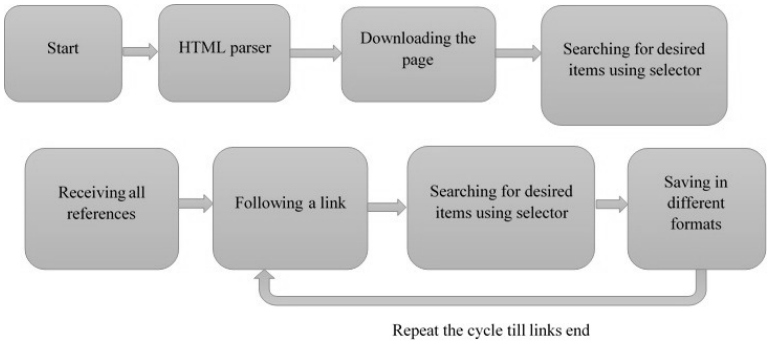

Figure 11: The scheme of parising and gathering information algorithm

of several input streams and allow to save content in different formats [13]. The results of parsing must be converted to the format which is suitable for further use. The exact format depends on the algorithm which is used to compare the information. In most cases XML and RSS-stream are most convenient, but result can also be saved in CSV-file, as this text format is very easy for further processing, easily converted to SQL-queries and Excel-compatible, or in XLS/XLSX format.

Here is the general scheme of the algorithm for detection of wrong and non-relevant information on websites. It compares the content of the webpage with the reference database. The alternative approach which can be used when no reference

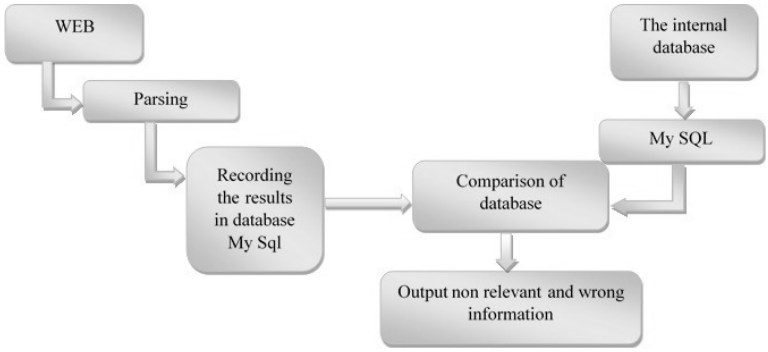

Figure 12: The architecture of the algorithm for detecting non-relevant and wrong information in websites.

database is available involves search and comparison of multiple sources of information in the Internet [14]. To identify the information which can be treated as reliable the mechanism of voting is used. The information provided by the majority of sources is considered to be reliable. When all three sources provide different response the number of sources should be increased until the majority of sources will provide the same information. It should be noted that sources of information should be independent. Therefore the algorithm must analyze the dependency between sources of information to avoid the situation where several sources of information are fed from the same source. Only when sources are independent voting mechanism can be used to identify the reliable information. The above algorithm was implemented and tested on the FCIT TNEU website. For the purpose of testing five incorrect changes were made to the teacher's positions on the FCIT TNEU website. 12 independent websites were used to identify the false information. The algorithm detected all 5 incorrect changes.

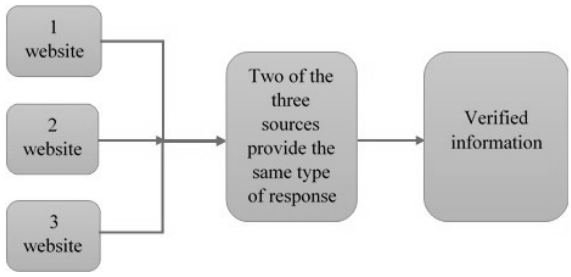

Figure 13: Voting algorithm using three sources

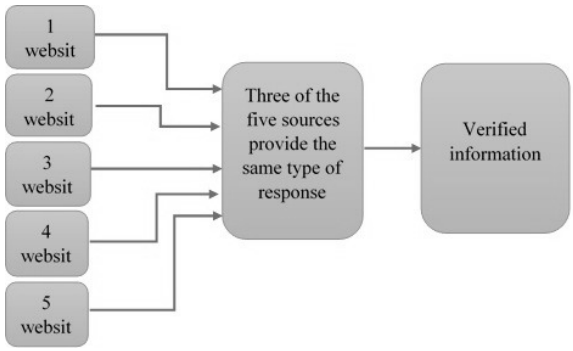

Figure 14: Voting algorithm using five sources

6. Conclusions

In this article the algorithm for automated detection of non-relevant or wrong information on websites was introduced. The algorithm extracts the semantic information from the webpage using available third party software and compares the semantic information with the reliable resources or with the content available in the Internet. In order to identify the reliable information in the Internet majority voting is used. The algorithm was implemented and tested on the FCIT TNEU website. All known non-relevant or wrong information was detected by the algorithm, which proves it's robustness and reliability.

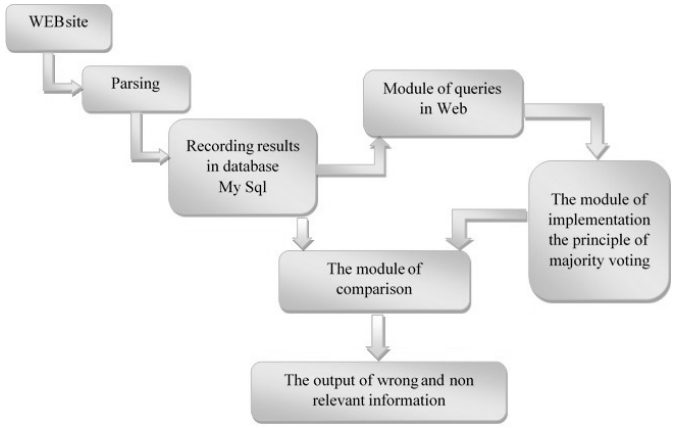

Figure 15: The architecture of the algorithm for reliable information identification using majority voting

References

[1] Pasichnyk, N. R. and Dyvak, M., *Formalism in the quality site creating problem*, Naukovi pratsi DonNTY, ser. Informatyka, kibernetyka ta obchysliuvalna tekhnika, Vol. 14, No. 188, 2011, pp. 325–329.

[2] Pasichnyk, N. R. and Dyvak, M., *Matrix the method and algorithm of construction of the content websites structures based on the ontological approach*, Naukovi pratsi DonNTY, ser. Informatyka, kibernetyka ta obchysliuvalna tekhnika, Vol. 15, 2012, pp. 184–189, (in Ukrainian).

[3] Pasichnyk, N., *Method of forming an ontological content, based on analysis of information at specialized Web-sites*, Visnyk HNU: Tekhnichni nauky, Vol. 5, 2012, pp. 241–244, (in Ukrainian).

[4] Pasichnyk, N., P. R. and Dyvak, M., *Mathematical model of traffic dynamics of the specialized websites and methods of its identification*, Induktyvne modeliuvannia skladnykh system: Zb. nauk. pr., Vol. 5, 2013, pp. 237–247, (in Ukrainian).

[5] Dyvak, M., P. R. and Pasichnyk, N., *Identification and modeling of limiting factors systems*, Proceedings of the 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), 2016, pp. 336–340.

[6] Dyvak, M. and Kowbasistyj, A., *Specific features of construction the method of detection the outdated and incorrect information on web resources*, Proceedings of the VI Ukrainian school-seminar for young scientists and students Advanced Computer Information Technologies, 2016, pp. 120–121.

[7] *The structure of the site. Creation and development of categorization*, url: http://seo-for-ucoz.com/load/podgotovka_k_prodvizheniyu/struktura_sajta/1-1-0-4 (in Russian).

[8] *Analysis of site structure*, url: http://www.web-patrol.net/audit-site-struktur.html (in Russian).

[9] *Information about HTTP status codes:*, url: http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html.

[10] *Information about User-Agent headers:*, url: http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html.

[11] *Information about the XPath language:*, url: http://www.w3schools.com/XPath/default.asp.

[12] Xin Wei, James Cai, J. R., *Use Base SAS URL to Build Surveillance and Monitoring System for New Clinical Trial Registration*, PharmaSUG 2010 Proceedings, 2010, url: http://www.pharmasug.org/cd/papers/AD/AD23.pdf.

[13] *Duncan Temple Lang. XML: Tools for parsing and generating XML within R and S-Plus*, url: http://CRAN.R-project.org/package=XML.

[14] *Duncan Temple Lang. RCurl: General network (HTTP/FTP/...) client interface for R*, url: http://CRAN.R-project.org/package=RCurl.