D. WILK-KOŁODZIEJCZYK, A. OPALIŃSKI, E. NAWARECKI, S. KLUSKA-NAWARECKA

ISSN 0543-5846
METABK 55(1) 127-130 (2016)
UDC – UDK 621.03:621.746:518.6.001=111

EXPLORATION OF WEB RESOURCES IN THE DOMAIN OF METAL PROCESSING TECHNOLOGIES

Received – Primljeno: 2015-01-27
Accepted – Prihvaćeno: 2015-07-30
Review Paper – Pregledni rad

The amount of information contained in the WEB grows in a galloping way, which is caused by the spread of Internet access and lowering the cost of storing and sharing data across the network. The vast amount of data, impossible to be analyzed by human, is the reason why finding and selecting valuable information has become a serious problem. Due to this situation, a highly useful and desired solution would be the development of a system that would allow continuous monitoring of the WEB and finding for the user valuable information from the selected Internet resources. This paper describes the concept of such a system, along with its initial implementation and application to search for information in the field of foundry industry.

Key words: technologies, casting, sectoral WEB monitoring, casting knowledge exploration

INTRODUCTION

Recent studies show that almost 2,4 milliard people have access to the Internet. In Europe it is more than 63 % and in North America more than 78 % of the population [1] [2]. Unfortunately, increasing the amount of data being in the network does not go hand in hand with the improvement of its quality. It causes the deterioration which in the literature is known as a phenomenon called "infobesity" [3]. In this situation, to find valuable information in the huge resources of the network is no small problem, requiring the use of specialized algorithms [4] [5]. On the market of products allowing search for information in the WEB resources, there are many general-purpose and specialized solutions. However, when it comes to searching for and monitoring of industry information from specialized industries, the offered functionalities often prove inadequate [6]. In response to this phenomenon it was decided to create a system that would allow the search for valuable information in the resources of the Internet, and monitoring changes in its content. The system provides the ability to define the precise patterns of information searched and of the mechanism sorting the results and making their presentation.

EXISTING SOLUTIONS

When it comes to the division in terms of functionalities offered and the data on which they operate, the existing solutions of information search in the WEB, can be divided into four main groups: universal search engines, page directories and WEB search (crawl) systems. Each group of solutions has different functionality, principle of operation and range of supported data.

A common feature of universal search engines is that theoretically they are free from any limits when it comes to the search for data. Search engines are assumed to be responsible for covering all the relevant WEB information resources, thus requiring huge financial outlays for equipment and infrastructure, which means that on the market there is only a small number of such solutions. The most popular universal search engines have many features in common. These include: the implicit algorithm for searching information and sorting the results found, and limit in the amount of available data [7]. Universal search engines allow creation of queries in a mode defined in advance by the administrators of the site. These features are the reason why the universal search engines may prove to be the tool insufficient in the case of search for sectoral industry information in the specialized WEB resources.

The second group of solutions, which can be useful when searching for sectoral industry information on the Internet, are directories of web pages that contain in their resources information and links to websites of companies, institutions, manufacturers, suppliers of materials, specific to a given industry. Most web directories have a built-in search engine, allowing search for specific items within their resources. However, the search is normally based on tags and keywords that have been assigned to an entry in the directory, and does not take into account the content provided in the source resources of the domain of a given entity.

The last type of tools that could be used in the WEB site search for information are the Internet searching systems (crawlers)[8] [9]. Here one can distinguish

D. Wilk-Kołodziejczyk, S. Kluska-Nawarecka, The Foundry Research Institute, A. Opaliński, E. Nawarecki, AGH University of Science and Technology, Krakow, Poland

both closed and commercial solutions as well as free and open systems. Examples of closed systems include Fast Search Web Crawler, which does not offer the possibility to expand and implement one's own data processing components found on the web, or Web Fountain which is a platform for analysis and processing of unstructured data, subscribed and licensed on a commercial basis [10]. The group of open systems includes projects such as Apache Nutch [11] and Datapark Search [12], allowing search and indexing of text data within the selected domains. Systems provide the ability to save and later search the web pages, but do not allow creating one's own components which could process the content of the page in the course of its search.

The need to be able to monitor and search for information in the resources of the Internet and the inadequacy of solutions available on the market were the reason for proposing the concept, design and subsequent implementation of a system meeting these functionalities.

SYSTEM CONCEPT AND ARCHITECTURE

The basic assumption was to create a system that would provide the user with the ability to monitor the Web sources and obtain the information valuable for him. The main functionalities of the system, which distinguish it from the existing solutions and make it useful for the user are:

- the ability to define a set of Internet domains, which are the subject of subsequent, periodically repeated search,
- the ability to define one's own patterns for searching, using the dedicated grammar,
- the ability to detect changes in the content of the pages containing the results found earlier.

A simplified system architecture is shown in Figure 1. It consists of four major components. The first of these components, a graphical user interface, provides the user with the ability to manage the system, define domains subject to search and monitoring, and define patterns of information that will be searched. With this component, the user also receives information about the results found, which may be freely viewed and analyzed. These functionalities are implemented in the system core component, containing all the business logic for the system operation, synchronization algorithms for domain search, access to the data in the database and to the subsequent search and analysis of results. Another component is a database that stores both the sources of the searched WEB resources and information about

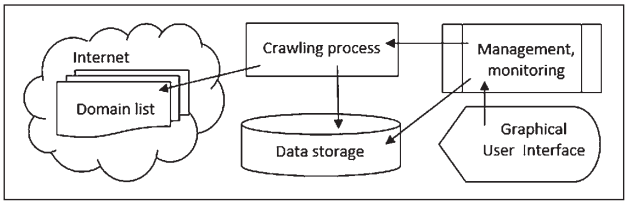

Figure 1 A system architecture schema

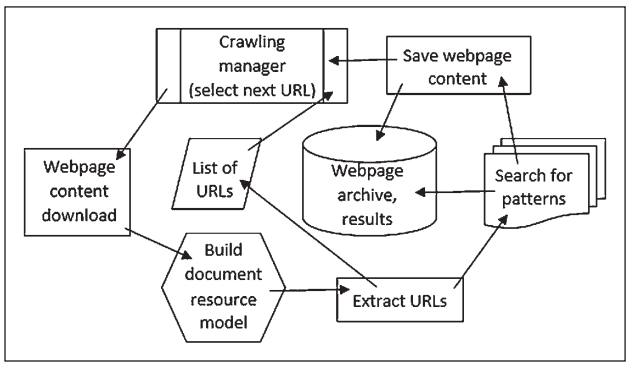

Figure 2 Stages of the crawling process

the patterns found in their content. All of these data are entered into the database by the crawl component, responsible for the process of searching the web domains and finding in their content patterns sought by the user. The steps of processing the data in the context of this component are given in Figure 2.

The crawling process begins by downloading a web page source for the previously selected URL. Based on the retrieved HTML code, a resource pattern is constructed for the page, and then the process of extraction of URLs found in the content of the processed page is undertaken. Then the advanced processing mechanisms for the Web page source are started, where one of them is composed of search algorithms for user-defined patterns. The next step is to save in the database both processed content page as well as all the additional information about the patterns found. The last step in the processing loop of resources is the choice by the crawling manager of another URL that will be processed in the next iteration of the loop.

A key aspect of the system operation is the ability to define patterns that will be searched during the operation of the system. Every single pattern in the system can be defined by a dedicated grammar. Thus developed grammar allows defining, in a sufficiently precise way, patterns that are next searched in the course of processing the WEB resources crawled by the system. The processing of each HTML page that is searched by the present system follows the dedicated algorithm The result of this algorithm is updating in the database information about patterns found on the crawled pages, described by grammars.

SYSTEM OPERATION AND TESTING

In order to verify the effectiveness of the designed system, an experiment was carried out. It consisted in using this system for searching and monitoring information from the foundry industry. Additionally, the same tests were carried out in parallel using the two most popular tools available on the market, i.e. the Google Search [13] engine and the Google Alerts [14] network monitoring system. The system presented in this study operates based on open sources of information from the Internet in Polish. There is the possibility

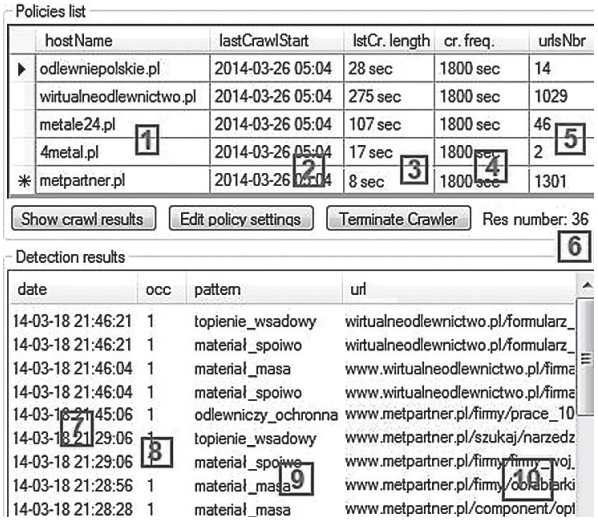

Figure 3 The application window, the results of searching for casting information

to use the system for data processing in other languages, but it is necessary to previously adapt to this task the natural language processing module (the question is mostly about the stemmer module searching for inflected forms of a particular word). For tests, five Internet domains containing information on metallurgical and foundry industry (data of companies and suppliers, news) were selected. These domains include: *metpartner.pl(A), 4metal.pl(B), metale24.pl(C), wirtualneodlewnictwo.pl(D), odlewniepolskie.pl(E)* (in parentheses are abbreviations of domains used later in the tables).

The next step was to nominate a set of patterns that were to be searched in the resources of previously selected domains.

It was decided to choose 22 different patterns described with grammar, based on combinations of 3 keywords from the set: *melting, charge material, cast iron, flux, ferroalloy, scrap, refiner, alloy, moulding, sand, binder, moulding mixture, coating, release agent, pattern, mould, core, shell, liquid, casting, filter, inoculant*. Specified grammar required the presence of all the three keywords within a single segment and allowed for various forms of inflection of the keywords considered in the pattern.

Figure 3 shows the results of the system operating within seven days, searching five domains previously selected once every 30 minutes. The results are ranked in order of their finding. The most important information available to the user shown in the figure is marked with numbers. The meanings are successively as follows:

1. The list of domains selected for monitoring. 2. The date of commencement of the last crawling process. 3. The duration of the last crawling process. 4. The domain searching frequency (in seconds). 5. The number of URLs found in the resources of the domain. 6. The total number of results (pattern matches) found by the system. 7. Date of finding pattern matches. 8. The number of changes in the content of the page or segment. 9. The abbreviated name of the pattern that has been found. 10. URL of the page containing the pattern.

Observations based on the progress and results of the search and monitoring of domains are as follows:

- Throughout the whole monitoring period 36 results were found, of which 34 were found in the first search of the domain; 2 others emerged on the 4th day of tests.
- The results were found only in two of the five domains selected for testing.
- 3 out of 36 results have changed during the monitoring process (the content of the segment under which they were found has changed).
- Adjustments were obtained only for 7 of the 22 patterns. The number of matches for each pattern is as follows (in parentheses the abbreviated names of patterns used in Table 1 are given):

In parallel, during testing of the presented system, comparative tests were carried out with the search tools offered by Google. To search the data, a universal Google Search engine was used. To monitor the network - Google Alerts tools were applied.

Search and monitoring were performed for all 22 phrases used in the operating system. To increase the transparency of the results, a comparison of the operation of the systems was presented in Table 1 for only 7 out of the 22 phrases, which were found by the system described in this article.

Table 1 **Performance of the presented system as compared with Google tools**

Pattern numbers	Number of pages matched		Number of observed changes	
	System	Google Search	System	Google Alerts
1	12	3	0	0
2	10	16	2	0
3	7	15	0	0
4	2	9	1	0
5	2	94	0	0
6	2	11	0	0
7	1	156	0	0

Rows of Table 1 correspond to the individual patterns, which have been found by the system. Columns 2 and 3 contain numbers of results found in domains where patterns were searched (2 domains by the presented system and 5 domains by the Google Search). Striking is the fact that Google Alerts tool did not return any result of new information appearing during the whole process of monitoring the network (column 4 in Table 1). Moreover, it should be noted that monitoring was not confined to only 5 domains, as in the case of the presented system, but the scope of the domains covered all websites. At the same time, the system presented, detected 3 changes in pages detected earlier (Column 3 in Table 1). This makes it a much more effective tool than the popular Google Alerts tool.

The main conclusions regarding the search process, which can be drawn on the basis of the results obtained, are as follows:

1. Viewing the html page sources with individual results returned by the Google Search engine it can be seen that in large part of the returned pages, at least one of the keywords is missing, which makes these results de facto mismatched to the model of grammar used in the compared system. Additionally, if in the content of the page all grammar keywords occur, they are traced throughout the whole site, and not only in the lowest segments, as it happens in the system presented in this article. The outcome is a larger number of results, but they are characterized by looser link with the subject.
2. Google Search mechanisms have greater opportunity to penetrate the site. Standard crawling algorithm of the presented system, only in 2 of the 5 sites exceeded the number of 1000 pages within a domain.
3. Some of the results found by the system presented in the article have not been found by Google Search.

Summarizing the results of the second part of the tests, which consist in comparing the operation of the presented system with the information search and monitoring tools offered by Google, the advantages of the newly developed system can be described as the following ones:

- higher efficiency in the process of data monitoring,
- better precision of information retrieval (real consideration of all the keywords and a precisely specified environment),
- the ability to define own methods for sorting the returned results.

Elements in which the presented system is inferior to the Google Search design are:

- limited number of domains that can be searched (lack of the possibility to monitor the entire Internet),
- limited possibility of penetration of some domains (simpler crawling algorithms).

Limitations of the presented system are, however, the result of hardware architecture on which the system is designed (one computer as opposed to the entire Google hardware infrastructure).

SUMMARY

The results presented in this article allow concluding that the proposed approach and its implementation meet the objectives with which they were created. The system allows the user to define his own patterns of queries and the extent of the domain and time range within which the source data are to be monitored.

The system already in its current version has demonstrated suitability for use in the field of foundry practice. It provides the possibility of being applied also in other industries, after changing the sets of patterns and the domain scope of the data monitoring. It also makes a promising platform for further development of information processing algorithms of which it is composed.

In short, the presented system is effective in finding accurate and high-quality information from online resources, eliminating the main drawbacks of the tools currently existing on the market. Additionally, it allows for continuous monitoring of resources and returning information about the newly-emerging content for search.

Acknowledgments

Work financed from funds of project nbr: 820/N-Czechy/2010/0.

REFERENCES

[1] S. Teltscher, Proceedings, Measuring the Information Society, International Telecommunication Union, A. Pitt, B. Granger (ed.), Place des Nations, Geneva, Switzerland, 2012, pp. 228-229.
[2] Miniwatts Marketing Group: World internet usage and population statistics, June 30, 2012, http://www.internetworldstats.com
[3] S.J. Bell, The infodiet: how libraries can offer an appetizing alternative to Google, The Chronicle of Higher Education, 50 (2004) 24, 15.
[4] S. Kluska-Nawarecka, B. Śnieżyński, W. Parada, M. Lustofin, D. Wilk-Kołodziejczyk, The use of LPR (logic of plausible reasoning) to obtain information on innovative casting technologies, Archives of Civil and Mechanical Engineering,14 (2014) 1, 25-31.
[5] S. Kluska-Nawarecka, K. Regulski, M. Krzyżak, G. Leśniak, M. Gurda, System of semantic integration of non-structuralized documents in natural language in the domain of metallurgy, Archives of Metallurgy and Materials, 58 (2013) 3, 927-930, DOI: 10.2478/amm-2013-0103
[6] A. Opaliński, W. Turek, K. Cetnarowicz, Proceedings, Scalable web monitoring system, Federated Conference on Computer Science and Information Systems (FedCSIS), Krakow, Poland, 2013, pp.1273-1279.
[7] I. Olejarczyk-Wożeńska, A. Adrian, H. Adrian, B. Mrzygłód, Parametric representation of TTT diagrams of ADI cast iron, Archives of Metallurgy and Materials, 57 (2012) 2, 613-617.
[8] M. Burner, Crawling towards eternity: Building an archive of the world wide web, Web Techniques Magazine, 2 (1997) 5, 37-40.
[9] C. Olston, M. Najork. Web crawling, Foundations and Trends in Information Retrieval, 4 (2010) 3, 175-246, DOI: 10.1561/1500000017.
[10] D. Gruhl, How to build a WebFountain: An architecture for very large-scale text analytics, IBM Systems Journal, 43 (2004) 1, 64-77, DOI: 10.1147/sj.431.0064
[11] R. Khare, D. Cutting, K. Sitaker, A. Rifkin, Nutch: A flexible and scalable open-source web search engine, Oregon State University, 1 (2004), 32.
[12] V. Hassler, Open source libraries for information retrieval, Software, IEEE, 22 (2005) 5, 78-82.
[13] www.google.com
[14] www.google.pl/alerts

Note: The responsible for English language is Krystyna Bany from Foundry Research Institute Krakow, Poland