ARCHIVES of FOUNDRY ENGINEERING

ISSN (1897-3310)
Volume 15
Special Issue
2/2015

43 – 46

Published quarterly as the organ of the Foundry Commission of the Polish Academy of Sciences

10/2

Domain WEB Monitoring

S. Kluska-Nawarecka [a], A. Opaliński [b], D. Wilk-Kołodziejczyk [a, *]
[a] Foundry Research Institute, Cracow, Poland;
[b] AGH University of Science and Technology, Cracow, Poland;
*Corresponding author. E-mail address: dorota.wilk@iod.krakow.pl

Received 09-042015; accepted in revised form 01.06.2015

Abstract

The last few years have seen a very dynamic development of the Internet worldwide. This is related to the rapid growth of the amount of information stored in its resources. The vast amount of data, impossible to be analyzed by man, is the reason why finding and selecting valuable information from a large number of results returned by search engines has recently become the task very difficult. Another problem is the low quality of the data contained in a large part of the results returned by search engines. This situation poses serious problems if one searches for detailed information related to the specific area of industry or science. In addition, the lack of effective solutions, allowing for continuous monitoring of WEB in terms of the search for emerging information while maintaining the high quality of the returned results, only aggravates this situation. Due to this state of affairs, a solution highly welcome would be a system allowing for continuous monitoring of the WEB and searching for valuable information from the selected Internet resources. This paper describes a concept of such a system along with its initial implementation and application to search for information in the foundry industry. The results of a prototype implementation of this system were presented, and plans for its further development and adaptation to other sectors of the industry were outlined.

Keywords: Web monitoring, Casting, Data integration

1. Introduction

The evolution of the Internet to WEB 2.0 model, where each user has the ability to create and place in a network his own content, has led to an avalanche growth in the amount of information placed on the network. Increasing availability of access to the Internet and reduced cost of these services contributed even more to this state. Recent studies show that almost 2.4 billion people have access to the Internet. In Europe it is more than 63% in North America 78% of the population [1], [2]. Unfortunately, increasing the amount of data being placed in the network does not go hand in hand with the improvement of its quality. On the contrary, it causes a deterioration, which in the literature is called "Infobesity" [3]. In this situation, to find valuable information in the huge resources of the network is no small problem, comparable to finding a needle in a haystack, requiring a huge amount of time, or the use of specialized algorithms [4]. On the market of products allowing search of the WEB resources, there are many general-purpose and specialized solutions. However, when it comes to searching for and monitoring of sectoral information from specialized industries, the offered functionalities often prove inadequate.

To find information in the WEB resources we can use a wide range of solutions offered, classifying them in terms of the distribution model used as dedicated systems, industry directories, and services such as SaaS (Software as a Service) offering the results of the operation of a system running at the service provider. When it comes to the division in terms of functionalities offered and the data on which they operate, the existing solutions can be divided into four main groups:

- universal search engines,
- pages of directories,
- service monitoring information in the WEB,
- WEB search (crawl) systems.

Each group has different functionality solutions, principle of operation and range of supported data.

A common feature of universal search engines is a theoretical lack of limits when it comes to the range of the data searched. Search engines are assumed to be responsible for covering all the relevant information resources in the WEB, requiring huge financial outlays for equipment and infrastructure, which means that on the market there is a scarce number of such solutions. Additionally, in the last decade, on the market of universal Internet search engines, we can observe a tendency to market monopolization by Google (www.google.com), leaving other search engines far behind, and gradually causing their marginalization. Google Search has now more than 71% of the network users, the second position is occupied by the Chinese search engine Baidu (www.baidu.com) with 16.5% of the market, while the 3rd and 4th place belong to Yahoo (www.yahoo.com) and Bing (www.bing.com), respectively, with the coverage of less than 6%. The overall results of other search engines fluctuate around 1% [5]. The most popular universal search engines have many features in common. These include: the implicit algorithm for information searching and sorting the results found, and limited amount of available data. For example, the Google Search engine for popular search terms, as the number of pages containing the required keywords often gives numbers up to several million, while the number of links available as search results rarely exceeds a few hundred. We do not know which results have been selected as worth presenting to the user and which have not been put on the list. Neither do we know on what basis the choice has been made [6].

Although universal search engines usually have the ability to add additional search options, such as logical operators, types of documents, or time range and domain, they only allow creation of queries in a mode as defined in advance by the administrators of the site. These features are the reason why the universal search tool may be insufficient in the case of the search for specialized industry information in the WEB resources.

The second group of solutions, which can be useful when searching for sectoral information on the Internet, are industry directories containing in their resources information and links to websites of companies, institutions, manufacturers, suppliers of materials, and other content specific to a given industry. Directories of domains can be either universal, covering many fields, to mention YellowPages (www.yellowpages.com), Open Directory Project (www.dmoz.org) and Best of the Web (www.botw.org), or focused on specific sectors and regions, such as the Industry Stock (www.industrystock.com) and Food Suplier (www.foodsupplier.com). 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. In addition, most web directories are commercial sites, and presentation of their content, the amount of contained information and the order of results returned usually depend on the appropriate option, which has to be purchased by individual companies and institutions located in the service resource. Another drawback of such solutions is small dynamics of the resource content, usually changing only by the administrator intervention, or update of entries introduced by users of the system [7]. This results in a mediocre representation of changes in the content of the directory pages with respect to possible changes in the landing pages regarding the described entities.

Another group of solutions, which might be useful in searching for sectoral information from the Internet, are network monitoring services offered by specialized companies operating in this field, such as CISION (www.cision.com) or NewtonMedia (www.newtonmedia.com). They offer services to monitor major news portals and user-defined domains, based on the searched terms and keyword sets. By the mere principle of operation, however, these are closed and paid solutions, and additionally, one of the stages of the information processing is human verification of the returned data, necessary to ensure its high quality.

The last type of tools that could be used in the search for information in the WEB are the Internet search (crawl) systems [8], [9]. Here one can distinguish both closed, commercial solutions as well as systems of free and open access. Closed systems include, for example, 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 Apache Nutch project [11], DataparkSearch [12] and Heritix [13], allowing search and indexing of text data within the selected domains. The systems provide the ability to save and later search the web pages, but give no chance to create own components which can process the content of the page, while it is being searched.

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, which meets these functionalities. The system is based on the platform of crawling and indexing information from the Web [14]. This article presents the results of the work on the development of this platform for monitoring and search for information regarding casting processes within a group of industry-defined domains.

2. The system concept and architecture

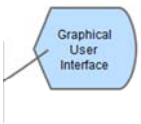

Fig. 1 A simplified system architecture

The basic assumption adopted in the implementation work was to create a system that will provide the user with the ability to

monitor the WEB sources for information valuable to him. The main functionalities of the system, which distinguish it from other existing solutions and make it useful for the user, are:

- the ability to define a set of Internet domains that are the subject of a subsequent search,
- the ability to define one's own patterns for searching using the dedicated grammar,
- the possibility of cyclic monitoring of selected resources,
- the ability to detect changes in the content of the pages containing the results found earlier,
- the ability to define one's own algorithm sorting the results returned by the system.

A simplified system architecture is shown in Figure 1. It consists of four major components. The first of these, a graphical user interface provides the user with the ability to manage the system, to define the domains subject to search and monitoring, and to define patterns of information that will be searched. With this component, the user also receives information about the results found and can freely view and analyze them. The latter functionalities are implemented in the next component, the core of the system, containing all the business logic for the system operation, algorithms for the synchronization of domain search, access to the data in the database, and 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 patterns found in their content. All these data are supplied to the database by crawl component, responsible for the process of searching the web domains and finding in their content patterns sought by the user.

3. Example of operation

In order to verify the effectiveness of the designed system, an experiment was carried out. The essence of the experiment was to use the developed system in searching for and monitoring of information from the foundry industry. The system operation is based on open sources of information from the Internet in Polish, but it is also possible to use it for data processing in other languages. However, for the latter purpose, the natural language processing module must be first adapted to this task (the problem is mostly the adaptation of stemmer module, which searches for inflection forms in selected words). For the tests were selected 5 Internet domains containing information (data on companies, suppliers, news) in the field of metallurgical and foundry industry. These domains are: *metpartner.pl, 4metal.pl, metale24.pl, wirtualneodlewnictwo.pl, odlewniepolskie.pl*.

The next step was to select 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 (K, 100%, segment 1), where the set K consisted of combinations of 3 keywords from the set: *melting, charge material, cast iron, flux, ferroalloy, scrap, refiner, alloy, moulding, sand, binder, moulding mixture, mould coating, release agent, pattern, mould, core, core coating, liquid, casting, filter, inoculant*. Parameter Tr = 100%, and L = segment assumed the necessity of the presence of all the 3 keywords in a single segment. The last parameter of the grammar V = 1 allowed for various forms of inflection to be considered in the standard keywords.

Fig. 2 The application window and the results of search for casting information

Figure 2 shows the results of the system operating within seven days and searching five domains previously selected once every 30 minutes. The results are arranged in the same sequence in which they were found. The most important information available to the user shown in the drawing is marked with numbers. Its meaning is successively the following:

1. The list of domains selected for monitoring.
2. The date of commencement of the last crawl process.
3. The duration of the last crawl process.
4. The frequency of searching the domain (in seconds).
5. The number of URLs found in the resources of a given domain.
6. The total number of results (matching the pattern) found by the system.
   Further information relates to the list of the search results found.
7. Date of finding the pattern matches.
8. The number of changes introduced to the page content 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 other occurred on the 4th day of tests.
- The results were found only in two of the five domains selected for testing.
- Service called *4metal.pl* rejected the possibility of search after downloading 2 pages from the site.
- 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 presented system provides two different modes to view the results found in the process of search and monitoring. The first mode is the direct upload of the website, on which the result was found. However, in the case of dynamically changing services, it may happen that the content of the page under the given address will change and will not contain any longer the data, which were

encountered during previous iterations of the service search. In this case, the user of the system is able to reproduce the page source with the results from the archives of the system, stored locally in the database (only text data)

Summary

The results presented in this paper allow us to conclude 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 time and domain within which the selected sources are to be monitored. This allows obtaining an adequate number of high-quality results, which can be analyzed by human in a further stage of processing. Thus we avoid the imperfections of other solutions available on the market: the black-box model and the lack of influence on the way the results are returned in the case of versatile search engines, combined with a low dynamics of the content in business directories.

The system even in its current embodiment has already demonstrated the suitability for use in the field of casting. It is also applicable in other branches of the industry after changing sets of patterns and the domain scope of the data monitored. It is also a promising platform for further development of information processing algorithms of which it is composed. Certain improvements are possible in the pattern component by introducing additional logical operators, as well as more advanced methods of sorting the results based on the degree of pattern matching. The plans for further development of the system anticipate expansion of the selection module of domains subject to monitoring. This would be done in the first phase of the system operation, based on the list of domains returned as search results using a standard universal search engine.

To sum up, the presented system is effective in finding accurate and high-quality information in online resources, eliminating the main drawbacks of the tools currently available on the market. Additionally, it also allows for continuous monitoring of resources and returning information about the newly-emerging searched content, which makes it very useful for the user in the difficult process of analyzing data from the Internet.

Acknowledgements

The work was financed within the framework of the international project No. 820/N-Czechy/2010/0 of 30 November and Financial support of the National Centre for Research and Development (LIDER/028/593/L-4/12/NCBR/2013).

References

[1] International Telecommunication Union: Measuring the Information Society 2012, Place des Nations, CH-1211 Geneva Switzerland, ISBN 978-92-61-14071-7.

[2] Miniwatts Marketing Group: World internet usage and population statistics (June 30, 2012), http://www. internetworldstats .com

[3] Bell, S. (2004). The infodiet: how libraries can offer an appetizing alternative to Google, *The Chronicle of Higher Education*. 50(24), B15.

[4] Regulski, K., Kluska-Nawarecka, S. & Wilk-Kołodziejczyk, D. (2015). Codification as a part of knowledge management in the research projects in the field of metallurgy. *Applied Mechanics and Materials*. 708, 288-293. DOI:10.4028/ www.scientific.net/AMM.708.288

[5] The Global Search & Social Report, Q1 2014, http://internationaldigitalhub.com/en/publications/the-webcertain-global-search-and-social-report-2014.

[6] Opalinski, A., Turek, W., Cetnarowicz, K. (2013). Scalable web monitoring system. In Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on, pp. 1273-1279.

[7] Chang, K.C.C., He, B., Li, C., Patel, M., & Zhang, Z. (2004). Structured databases on the web: Observations and implications. *ACM SIGMOD Record*. 33(3), 61-70.

[8] Burner, M. (1997) Crawling towards eternity: Building an archive of the world wide web. *Web Techniques Magazine*. *2*(5).

[9] Olston, Ch. & Najork, M. (2010). Web Crawling. *Foundations and Trends in Information Retrieval*. 4(3), 175-246.

[10] D. Gruhl et al. (2004) How to build a WebFountain: An Architecture for very large-scale text analytics. *IBM System Journal*. 43(1), 64-77.

[11] Khare, R., Cutting, D., Sitaker, K., & Rifkin, A. (2004). Nutch: A flexible and scalable open-source web search engine. *Oregon State University*. 1, 32-32.

[12] Vesna, H. (2005) Open source libraries for information retrieval. *IEEE Software*. 22(5), 78-82.

[13] Mohr, G., Stack, M., Ranitovic, I., Avery, D., & Kimpton, M. (2004) An Introduction to heritrix. An open source archival quality web crawler. 4th International Web Archiving Workshop.

[14] Turek, W., Opalinski, A., & Kisiel-Dorohinicki, M. (2011). Extensible web crawler–towards multimedia material analysis. In Multimedia Communications, Services and Security, CCIS, vol. 149, pp. 183-190. Berlin: Springer Heidelberg.