www.czasopisma.pan.pl PAN www.journals.pan.pl

ARCHIVES of FOUNDRY ENGINEERING

ISSN (2299-2944)
Volume 18
Issue 1/2018

212 – 216

DOI: 10.24425/118839

38/1

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

Investigation of the Process of Data Acquisition Regarding the Process of Casting from the Internet

D. Wilk-Kołodziejczyk [a, b] *, S. Kluska-Nawarecka [b], Z. Stefański [b], A. Smolarek [a]

[a] AGH – University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland

[b] Foundry Research Institute, ul. Zakopiańska 73, 30-418 Kraków, Poland

* Corresponding author. E-mail address: wilk.kolodziejczyk@gmail.com

Received 13.09.2017; accepted in revised form 02.02.2018

Abstract

Access to up-to-date information on technology, innovation, source publications, and the materials and services offered in a particular industry is very important for both industrial plants and departmental research centres. It should be noted that obtaining such information using publicly available search engines such as Google, Yahoo!, Bing, Bindu (mainly used in China) is only apparently easy because, due to their versatility, they deliver results with great redundancy. This leads to the need to analyze large data sets by linguistic methods or "manually", which is very tedious and time consuming.

In this situation, it was considered reasonable to undertake studies aimed at acquiring relatively simple IT tools, i.e. crawlers, which allow their users to selectively search for information in a particular problem area, which in this particular case is casting. The intention of this work was to collect and analyze the experimental material that would allow describing the characteristics of the above solutions from the point of view of the range of their application, the quality of the results achieved, and possible limitations and preferences taking into account user needs [1, 2].

Keywords: Up-to-date information technology, Crawler, Foundry industry, Casting process

1. Introduction

The problem of obtaining textual information from open sources, and in particular from the Internet, has recently become a subject of widespread interest aroused, on the one hand, by the ever-increasing role of the Internet as a global source of information and forms of the exchange of this information and, on the other hand, by the development of more sophisticated IT tools for the interpretation and analysis of data contained in text documents. It should be noted, however, that textual analysis based on the use of linguistic methods, although often leading to spectacular results, is characterized by enormous computational complexity, and its effective use demands a fairly broad knowledge from the user. For these reasons, the use of this class of methods under industrial conditions seems to be lacking any development prospects. As deep web grows at a very fast pace, there has been increased interest in techniques that help efficiently locate deep-web interfaces. However, due to the large volume of web resources and the dynamic nature of deep web, achieving wide coverage and high efficiency is a challenging issue [3, 4].

Web crawling is an important method for collecting data and keeping up to date with the rapidly expanding Internet. A web crawler is a program, which automatically traverses the web by downloading documents and following links from page to page. It is a tool for the search engines and other information seekers to

gather data for indexing and to enable them to keep their databases up to date [5]. All search engines internally use web crawlers to keep the copies of data a fresh [6]. A web crawler is a software or programmed script that browses the World Wide Web in a systematic, automated manner. Internet search engines use web-indexing software to update the content of websites or web content indexes of other websites. Web crawlers copy pages to be processed by the search engine, which indexes the downloaded pages so that users can search more efficiently [7].

Under previously implemented projects [8], two of such solutions, described in this study as "Dimension" and "Opal", were acquired.

It should be mentioned that the initial intention was to have a much larger scope of work, including also the identification of procedures for using the acquired information to create knowledge modules with specific content. However, due to the limitations imposed by both implementation time and financing options, it was decided to carry out actions which would directly concern the conditions of application of the examined tools and interpretation of the obtained results.

2. Characteristics of the examined solutions

The Crawler denoted by the "Opal" acronym is a prototype solution (tool) that is intended to be used to experimentally investigate factors with some impact on the effects of the Internet exploration process, taking into account especially sources related to foundry problems [9, 10].

The input data necessary for proper operation of the system are:

- list of terms to be searched
- list of domains that will be searched for terms.

The list of terms searched by the system, determined after consultations with the technologists, contains the following sets of keywords:

1. " melting", "cast iron", "flux",
2. " melting", "ferroalloy", "scrap",
3. " melting", " charge", " cast iron",
4. " melting", "refiner", "scrap",
5. " melting", "alloy", " refiner",
6. "moulding process", "moulding material", "sand",
7. " moulding", "casting", "pattern",
8. " moulding", "moulds", "cores",
9. " casting", "molten", "flux",
10. " casting", " molten", "refiner",
11. " casting", "treatment", " molten",
12. " casting", "material", "filter",
13. " casting", " molten", "inoculant",

The selected list of domains that could be potential search sources with comments about their content is as follows:

www.industrystock.pl - product and company directory, high dynamics
www.europages.pl - business directory, very high dynamics
www.pkt.pl - catalogue of companies, very high dynamics and large scope
www.123zapytanie.pl - announcements and inquiries, various areas, high dynamics
www.pl.all.biz - business directory, wide range
www.st3.pl - **good catalogue of foundry / sectoral companies, without keywords** (useful for generating source list)
www.oig.com.pl - little information within the domain, lots of links and info in pdfs
www.odlewnictwo.com - to buy
www.odlewnictwo.com/odlewnictwo/odlewnictwo_informator.htm - about 20 companies entered "manually"
www.czechtrade.co.uk/catalogues-exporters/ - a small but businesslike catalogue of Czech exporters with a brief Polish description of the company and a link to the site (usually in English)
www.odlewniepolskie.pl - interesting site, moderate dynamics (good for processing), contains information about materials, branches and offered products
www.wirtualneodlewnictwo.pl - good database of foundry companies (about 100)
www.metale24.pl - metallurgical portal, interesting, updated once a week, companies within the domain, moderate dynamics - good for crawling, poor transfer
www.panoramafirm.pl - catalogue of companies from all sectors
www.4metal.pl - **interesting site with metallurgical information, moderate dynamics, probably useful**
www.baza-firm.com.pl - database of companies, catalogues, various domains
www.katalog.wp.pl - same as above, very general
www.metpartner.pl - **catalogue of companies, products, ads, theoretically metallurgical, practically quite broad (analysis, protection)**
www.odlewnictwo.org - list and referrals for companies, well designed hierarchy but not in domain
www.aaazapytanie.pl - queries, general, too broad and dynamic
www.euro-info.com - company database, search engine, general
(bold letters mark domain names, which initially due to their characteristics were selected for search).

Even the initial series of tests has shown that search efficiency depends on the number of terms taken and on the number of keywords contained in them, as well as on the number and the choice of domains that are being explored. Based on the results obtained, it was decided to limit the number of keywords to 3 (as stated above), while the number of domains initially limited to 5 was later limited to three.

Sample results obtained for the 5 domains are shown in Figure 1. These are the following domains:

- metpartner.pl
- 4metal.pl
- metale24.pl
- wirtualneodlewnictwo.pl
- odlewniepolskie.pl

The domain crawl rate (the domain crawl frequency) was set to 30 minutes.

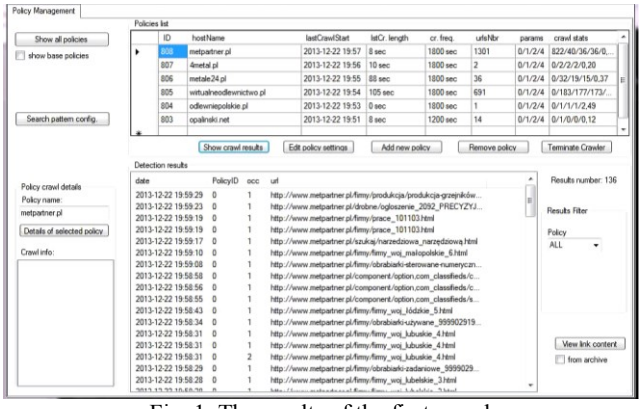

Fig. 1. The results of the first search

After eliminating the less effective domains, i.e. metapartner.pl and odlewniepolskie.pl, the search was repeated and the results shown in Figure 2 were obtained:

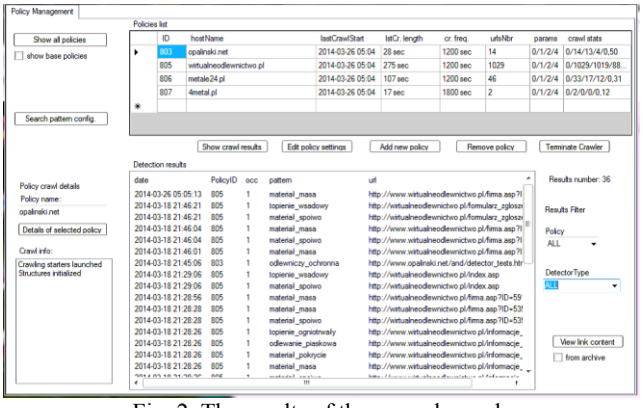

Fig. 2. The results of the second search.

For the set of domains such as shown in Figure 2, the system after a few minutes found 36 pages, the content of which corresponded to the assumptions adopted.

Figure 2 shows in the top part of the window *the names of the domains*, *the start date of the last crawl*, *its duration, the crawl rate of the domain, the number of URLs found in the domain resources, the search times and the assumed frequency*, and in the bottom part of the window - *the date of discovery, the number of page content changes, the searched terms* (the first and the last word of the pattern) and *addresses of the searched pages* (containing the pattern). The right part of the window contains *the number of results* (the total number of results found by the system).

In the case under discussion, the total number of 36 pages found will be broken down into the following terms:

- material, moulding, binder-12 results
- melting, material, charge-10 results
- material, moulding, sand mixture-7 results
- material, moulding, coating-2 results
- casting, mould, sand-2 results
- melting, material, refractory-2 results
- casting, coating, protective-1 result

The "Dimension" system is the second solution that has been used to conduct comparative research. *Dimension* is a universal IT tool for searching information from the Internet without limiting the problem area in which the search is to be conducted. Unlike the "Opal" system, it does not have an *a priori* defined set of terms that underpin the search, or the limitations or preferences for domain selection. In this system, for a specific task, a set of keywords (possibly in the form of regular expressions) is created, from which patterns (terms) can be built. They may contain 2, 3 or more elements, depending on the specification of the document sought.

As a consequence, the comparative studies of these two tools should be considered interesting and important from a practical point of view, as they can provide important guidance on the preferences and conditions of use.

In the description presented here, details about the form of windows used in the Dimension system and presentation of the search results are omitted as they are similar to the solutions already described in the Opal system. As a result, the number and URLs of pages with the pattern searched are also obtained, but it was considered crucial to carry out experiments to compare these two solutions and formulate conclusions about their preferences and the range of application.

3. Comparative studies

Comparison of the performance of the two systems that differ in their operating algorithms, internal structure, and procedures for the evaluation of solutions is not a trivial task, and can be done in a variety of ways.

In the adopted approach it has been decided to conduct a series of experiments that will allow numerical evaluation of the results obtained for the random selection of domains and the assumed forms of terms used in their search.

The obtained results of this analysis are presented in Table 1.

Table 1.
The number of results obtained from various domain searches

	Ferroterm.pl			***dobryodlew.pl***			***odlewniawarszawa.pl***			***hutazabrze.pl***			***lisiekaty.pl***			***pedmo.eu***			***odlewnia-zremb.pl***			***Google search***
	Number of pages found			Number of pages found			Number of pages found			Number of pages found			Number of pages found			Number of pages found			Number of pages found			Approximate number of pages found
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	
Term	Opal	W-wa	site:Name of domain	Opal	W-wa	site:Name of domain	Opal	W-wa	site:Name of domain	Opal	W-wa	site:Name of domain	Opal	W-wa	site:Name of domain	Opal	W-wa	site:Name of domain	Opal	W-wa	site:Name of domain	
G	**0**	**0**	**0**	**0**	**2**	**3**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**49 000**
I	**1**	**1**	**1**	**0**	**5**	**12**	**0**	**0**	**0**	**1**	**2**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**		**1**	**31 900**
M	**0**	**0**	**1**	**0**	**0**	**4**	**4**	**2**	**9**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**10 500**
N	**0**	**0**	**0**	**0**	**3**	**8**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**1**	**6**	**0**	**0**	**0**	**0**	**1**	**1**	**55 000**
Y	**0**	**0**	**1**	**0**	**1**	**2**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**8 670**
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**675**
H	**0**	**0**	**0**	**0**	**1**	**4**	**0**		**0**	**0**		**1**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**13 200**
J	**0**	**0**	**0**	**0**	**3**	**8**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**13 200**
O	**0**	**0**	**0**	**0**	**3**	**5**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**1**	**0**	**0**	**0**	**29 400**
T	**0**	**2**	**9**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**156 000**

Terms:
G - moulding process, moulding, sand
I – material, moulding, moulding sand
M - casting, mould, sand
N - moulding, moulds, cores
Y - melting, equipment, cast iron
F - melting, alloy, refiner
J - material, moulding, coating
H - material, moulding, binder
O - casting, coating, protective
T - casting, material, filter

The experiment was conducted for 7 domains and 10 terms determining the searched documents. For technical reasons, the experiments were divided into two groups - each consisting of a 5-term search (the fragment of Table 1 containing G, I, M, N and Y terms is written on a grey background).

To create a certain reference to the capabilities of searching for this type of results using publicly available search engines, the last column of the table gives the number of pages indicated by the Google search engine. It is clear from the numerical values given that the result of such a dimension is practically useless. On the other hand, the result in the "site: domain name" column is comparable to the "Opal" and "Dimension" crawlers, but it is simply "manual" crawling of the domain and "manual" checking, e.g. if in a week the result has changed and what possibly has changed. Both crawlers perform and archive this automatically.
In a similar way, one can perform experiments that include other sets of domains being searched and other terms according to which these domains are searched. Such sets may be based on the experience gained and user preferences for the selected problem areas.

As we can see, the data obtained in the form such as shown in Table 1 can be used to evaluate the search efficiency with respect to the domain, term set, and finally the tool used (here this tool is "Opal" or "Dimension"). This allows rationally choosing the searched domains, as well as the terms used. It is also possible to formulate conclusions about the utility of the tool (crawler).

4. Conclusions

The work carried out allowed for a more detailed examination of the specifics of operation of Opal and Dimension systems held by the Foundry Research Institute. The use of these tools to explore WEB resources has proved to be effective, avoiding the inconveniences of using publicly available search engines (Google, Yahoo! and others), inconvenient mainly due to redundancy and the difficulty of directing search to a specific problem area.

The tested solutions have mechanisms that allow automatic "refreshing" and archiving of possible changes occurring on the selected pages at given observation periods.
The relationship between the applications in question is as follows:

- Dimension is a universal system - it can be used in any problem area, it has no restrictions on the terms searched (number of keywords, semantics) or on the number of domains explored;
- Opal is an application dedicated to finding information about foundry industry as expressed in the definition of a set of terms, which is basically left unmodified during the operation of the system, while the definition of a set of domains belongs to the initial conditions of the task;
- from a practical point of view, the above differences result in a slightly shorter turnaround time for the Opal system, but the advantage of Dimension's application is the number of positively recognized sites;
- both systems, due to their relatively simple operation and unqualified hardware requirements, can be used in industrial and laboratory environments.

An important aspect of the work carried out was to investigate the impact of domain selection and term set on the search results. It has been found that different domains had different "susceptibility" to the exploration activities, but evaluation of this property can only be statistical.

An interesting idea for gaining numerical domain characteristics in the context of the terms used seems to be the approach shown in Table 1. Carrying out several series of experiments according to the procedure outlined there allows for a rational selection of the domains being tested and assigning to them appropriately selected terms.

Summing up the assessment of the studies carried out it can be concluded that they have led to a number of insights that will allow for more efficient web search procedures and formulating rules for the use of Opal and Dimension systems in industrial environments.

References

[1] Regulski, K., Kluska-Nawarecka, S. (2012). *Knowledge integration computer tools and algorithms in the improvement of the production processes of cast-steel castings, Artificial Intelligence in the Knowledge and Information Systems*. Kraków: Instytut Odlewnictwa.

[2] Kluska-Nawarecka, S., Wilk-Kołodziejczyk, D., Regulski, K. (2013). Formalisms and Tools for Knowledge Integration Using Relational Databases. In Nguyen N.T. (eds), Transactions on Computational Collective Intelligence XII. Lecture Notes in Computer Science, vol 8240. (pp.1-20) Berlin. Springer, Heidelberg Breiman.

[3] Kolias, V., Anagnostopoulos, I., Zeadally, S., (2017). Structural analysis and classification of search interfaces for the deep web. *The* Computer Journal. DOI: https://doi.org/10.1093/comjnl/bxx098.

[4] Zhao, F., Zhou, J, Nie, C., Huang, H., (2016). *SmartCrawler: A Two-Stage Crawler for Efficiently Harvesting Deep-Web* Interfaces, IEEE Transactions on Services Computing, vol. 99.

[5] Shetty, K.S., Bhat, S., Singh, S., (2012) *Symbolic Verification of Web Crawler Functionality and Its Properties,* International Conference on Computer Communication and Informatics (ICCCI-2012). Coimbatore, INDIA, IEEE Conference Publications.

[6] Udapure, T.V., Kale, R. & Dharmik, R. (2014). Study of web crawler and its different types. *IOSR Journal of Computer Engineering (IOSR-JCE)*. 16.

[7] Kausar, Md. Abu, Dhaka, V.S. & Kumar, S. (2013). Web Crawler: A Review. *International Journal of Computer Applications*. (0975-8887), 63(2).

[8] Final report on the implementation of the international non-financed project for the years 2014-2015, Developing solutions for conceptualisation and sharing knowledge about the components of casting technologies in the context of innovation and improvement of production processes (in Polish).

[9] Kluska-Nawarecka, S., Jančíková, Z., David, J., Wilk-Kołodziejczyk, D., Regulski, K., Dajda, J. (2014). Knowledge components description, support to prevent defects of metal products using methods based on artificial intelligence and ETL technologies. In 23rd International Conference on Metallurgy and Materials Metal 2014, 21-23 May 2014. Brno, Czech Republic.

[10] Opaliński, A., Nawarecki, E., Kluska-Nawarecka, S., (2015). Agent-based approach to web exploration process. *Procedia Computer Science.* 51, 1052-1061.
DOI: https://doi.org/10.1016/j.procs.2015.05.263.