CONFIGURATION MODEL FOR FOCUSED CRAWLERS IN TECHNOLOGY INTELLIGENCE

GÜNTHER SCHUH
Fraunhofer Institute for Production Technology IPT, Germany
guenther.schuh@ipt.fraunhofer.de

ANDRÉ BRÄKLING
Fraunhofer Institute for Production Technology IPT, Germany
andre.braekling@ipt.fraunhofer.de (Corresponding)

TONI DRESCHER
KEX Knowledge Exchange AG, Germany
info@kex-ag.com

ABSTRACT

Due to a steady increase of competitive constraints caused by ongoing globalization and dynamically growing markets, technology intelligence has become an important element of strategic business intelligence. The objective of technology intelligence is to focus on the systematic identification of future chances but also threats to companies caused by new technologies and further technology developments.

To operate technology intelligence efficiently, access to up-to-date, relevant, and sufficiently complete information is essential. Indeed, availability of information is higher than ever by reason of digitalization. However, it also causes the problem of information overload. The available mass of data has to be searched, assorted and assessed to identify the actual needed information. In addition, the entire information processing has to be continued permanently or to be repeated for each new object of investigation, otherwise the validity of the results is not given any more.

Accordingly, it appears reasonable to automate this process by widely using smart software solutions. One of the promising approaches is “focused crawling” which not just runs through given data sources in the web, but also rates each data record to make an autonomous decision, which information is relevant for the further process, and which data records should reasonably be analyzed next.

To implement such crawlers, different approaches exist in the field of information retrieval: For example, different rating and discovery algorithms. This paper presents the status quo of ongoing research to develop a configuration model for focused crawlers to fulfill the varying requirements of technology intelligence tasks. At first, the assessment criteria for information in a technology intelligence process and the configuration possibilities of focused crawlers are described. As a result, a first approach of a matching between the requirements of technology intelligence tasks and the consequences of different focused crawler configurations is presented. Closing, the paper explains how this approach will be improved and validated in case studies prospectively.

Key words: Technology Intelligence, Focused Crawler, Information Overload, Smart Software

INTRODUCTION

Due to the ongoing globalization, dynamically growing markets, and the emergence of new technologies, competitive opportunities and constraints are increasing rapidly. Derived from this, the necessity of an effectively and efficiently strategic business intelligence process is made obvious. Related to technology ventures, especially technology intelligence is indispensable to systematically identify prospective chances but also threats by new technologies and technology developments, as explicated by Wellensiek (2011).

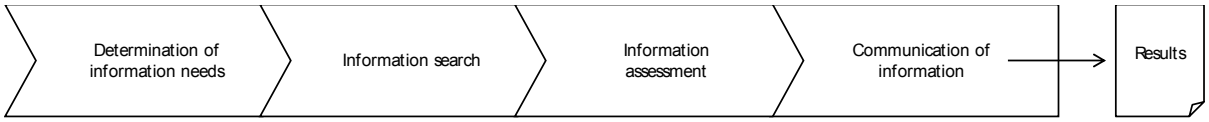

Figure 1: The four steps of the technology intelligence process according to Wellensiek (2011).

The process of technology intelligence is divided into four steps, as shown in figure 1. First, the information needs have to be determined. They are affected by the conducting company's business and technology strategy. During the step of information search, as much relevant information as possible related to the current information need is brought together using different information sources. On this basis, a further information analysis is performed. Finally, the results of the analysis are prepared to be communicated, e.g., as a final report.

Because of the digitalization and networking in today's information age, information is available easily. At first glance, this seems to be an advantage for technology intelligence. However, this also causes the problem of information overload. The following example shows the dramatic increase of data: Until beginning of 2012, the US Library of Congress collected data in the amount of 374 terabytes (US Library of Congress, 2013). This is about 73.5 million times the complete works of William Shakespeare[1]. The International Data Corporation (IDC, 2014) estimated the digital universe in 2013 as 4.4 zettabytes, and forecasts it will reach 44 zettabyte in 2020. This amount of data is obviously not manageable manually anymore. Therefore, the practical challenge is to make these data useful to technology intelligence by computer-aided research.

Focused Crawlers provide such an opportunity to support research processes. They are a smarter implementation of a web crawler. A simple web crawler collects all available documents in connected sources (e.g., the internet) by extracting and following all references in these documents. Whereas a focused crawler only considers those documents in its sources, that are relevant in the given research topic, as described by Chakrabarti et al. (1999).

The necessary relevance assessments are possible by different statistical and linguistic algorithms. In addition, Batsakis et al. (2009) and Pant/Srinivasan (2005) point out that the software requires an appropriate training related to the specific topic. The effectivity and the efficiency of a focused crawler depend on the selection of its components in terms of adequate algorithm implementations and training data sets related to the current topic and task. Some of the basic concepts and approaches of focused crawlers are described in Micarelli and Fabio (2007).

[1] Based on the UTF-8 text file "The Complete Works of William Shakespeare", which is available at http://www.gutenberg.org/ebooks/100.

According to Srinivasan et al. (2005), the software's effectivity is measured by its recall ratio and its precision, the efficiency is measured by the caused costs in the form of resources (CPU time, memory requirements). On the other hand, the requirements are defined by the different basic activities of technology management (scanning, monitoring, and scouting) as well as by the determined information needs.

Currently, there are many implementations of focused crawlers intended for specific use cases. To use these implementations in other contexts, usually components of their configuration have to be changed, e.g., new training data has to be defined or even algorithms have to be replaced. Therefore, focused crawlers are not easy and quick to use.

In this article we present our ongoing research project about the usage of focused crawlers in technology intelligence processes. Due to the mentioned constraints, the target of this project is a configuration model to support the configuration of focused crawlers by choosing appropriate components that fit the different requirements of technology intelligence tasks.

APPLICATION OF FOCUSED CRAWLERS IN TECHNOLOGY INTELLIGENCE

As mentioned before, the requirements for the usage of focused crawlers in technology intelligence are defined by the different basic activities scanning, monitoring, and scouting. Wellensiek et al. (2011) describe these activities as follows:

i. Technology scanning is an undirected but constant search process. It aims to give an overview about currently unknown but potentially relevant technology information. E.g., a company can scan trending topics like industry 4.0, big data or additive manufacturing independently of their current relevance to identify new opportunities and constraints in time.

ii. Technology monitoring is a long term activity to observe a previously identified relevant field of technology information (e.g., by a technology scanning) and performs a more specific search for new information related to this field. So a company can monitor a new technology that is very interesting but currently too expensive in production to become an early adopter if the manufacturing cost drops.

iii. Technology scouting aims to get fast and detailed information about a very specific technology by order. A scouting is done by a company if an explicit decision has to be made, e.g., about a market entry or an investment.

Each of these basic activities performs the four steps of technology intelligence processes, as explained in Spath et al. (2010) and also shown in figure 1:

i. Determination of information needs: First, a search field strategy has to be defined as explicated in Schuh et al. (2009). The search fields depend on the company's core capabilities and also on its business and technology strategy. They can be used as orientation guide, e.g. to specify the fields of observation, the required level of detail, and the used methods. Lichtenthaler (2005) examined the selection of technology intelligence methods in 26 large technology-intensive companies in Europe and North America.

ii. Information search: Based on the defined information needs, feasible sources of information have to be chosen, the necessary amount of data has to be stated and the actual research is

performed. Sources of information can for instance be patents, publications, and websites, but also conferences, fairs, expert interviews and discussions.

iii. Information assessment: Afterwards, the collected information has to be rated by its relevance for the current activity to extract a reasonable and manageable subset. This subset has to be considered and interpreted to deduce the process's key findings, e.g., about a monitored technologies maturity.

iv. Communication of information: Concluding, the results have to be prepared in an appropriate form for its audience. So they can be used in a management summary to support upcoming decisions, but also to define a new, more specific research process, e.g., if a scanning process identified possible opportunities before.

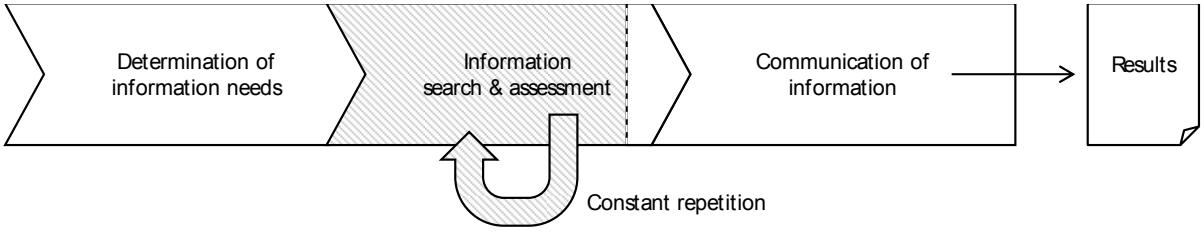

Figure 2:Technology intelligence process supported by a focused crawler (hatched area).

The assumption is, that a focused crawler is useful in the steps of information search and assessment in a combined and repeated form (see figure 2), because the focused crawler rates its findings already during the ongoing research process, especially, to decide which references to follow next, as described in Rawat/Patil (2013). Furthermore, the repetition provides a constantly updated database with manageable effort. Just the final consideration and interpretation has still to be done manually, of course. But to get meaningful results, it is essential that the crawler is well-configured in terms of the selection of components that meet the search field's requirements. These components and their implied correlation to the technology intelligence processes are described in the following paragraph.

ASPECTS OF A CONFIGURATION MODEL FOR FOCUSED CRAWLERS IN TECHNOLOGY INTELLIGENCE

To define the configuration model, factors and requirements related to both the process of technology intelligence and focused crawlers have to be identified. For that purpose, a basic focused crawler infrastructure is introduced subsequently before possible connections and resulting factors are discussed.

Infrastructure of Focused Crawlers

Figure 3 shows an exemplary focused crawler infrastructure which consists of 4 layers, see Pant/Srinivasan (2005). Based on the information needs, training data and the crawler configuration will be defined as input data. Afterwards, the data will be analyzed in the *intelligence layer* and a decision is made, which references will be followed next. These references (e.g., URLs) will be deposited in the so-called *frontier*, wherefrom the *network layer* will get and follow one after another, so the discovered contents can be stored in the *repository*. The *parsing and extraction layer* extracts all necessary information from the retrieved contents, before the *representation layer*

converts them into a reasonable and machine readable form. This form is the basis for the next analysis by the intelligence layer, whereupon the process begins again.

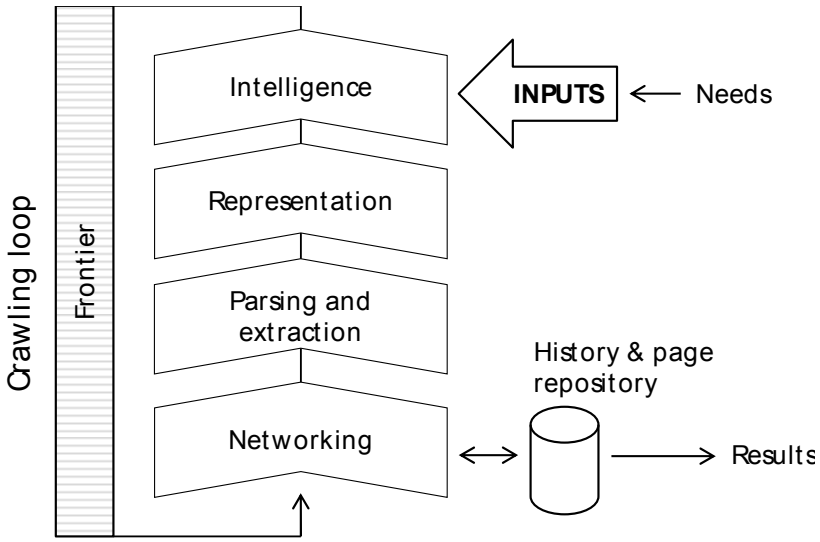

Figure 3: A focused crawling infrastructure according to Pant/Srinivasan (2005) and Schuh et al. (2014).

All these layers can be implemented by different, in parts also combined algorithms. Moreover, the selection and preparation of the input data has a crucial impact on the results of the crawling process, compare Rawat/Patil (2013) and Srinivasan et al. (2005). We call all these interchangeable parts *components* which should be composed by our intended configuration model. A further explanation of a focused crawler's mode of operation is (amongst others) described in Pant/Srinivasan (2005), Zhuang et al. (2005), Pal et al. (2009) and Singh/Tyjagi (2013).

Requirements towards Focused Crawlers

According to Srinivasan et al. (2005) the results of a focused crawling are evaluated by recall, precision, and cost. Precision describes the fraction of retrieved results that are relevant, and recall describes the fraction of relevant results that are actually retrieved, as stated by Manning et al. (2008, pp. 142-143). Cost includes parameters like CPU load, memory usage, data storage, and similar parameters. Schuh et al. (2014) describe recall, precision, and cost as interdependent: A better recall requires unavoidably more memory, and low computing power can be compensated by reducing the precision. As well, a high recall can cause a low precision, because strict relevance conditions may result in so called false negatives (i.e., relevant documents which are rated as irrelevant by mistake).

In consideration of this interdependence, the described evaluation parameters can be used to define a specific task's requirements towards focused crawlers.

Information Needs and Information Search

Corresponding to Wellensiek et al. (2011), at first the search field has to be stated to describe the actual information needs, because it has an appropriate impact on the subsequent research process. Schuh et al. (2009) enumerate following parameters for company-specific information needs: the basic characteristic of the company, its technological base, the technological and scientific environment and the available offer of information.

Before the actual research within the defined search field can start, the information sources have to be chosen, also according to Wellensiek at al. (2011). Hence, it is necessary to prove if and how these sources can be accessed. Also, the level of information detail has to be defined matching the processed activity and its requirements. A comprehensive network configuration model to set up an optimized technology intelligence network by integration of specific information sources is developed by Saxler (2011).

Requirements towards Technology Intelligence Processes

Schuh et al. (2009) mention information content, earliness, exclusiveness, and the information's usefulness as base for decision making and thus as requirements towards technology intelligence processes. Regarding to the company's technology strategy, these requirements have different specifications. E.g., a company which wants to be the technology pioneer in its target market needs access to early and exclusive information.

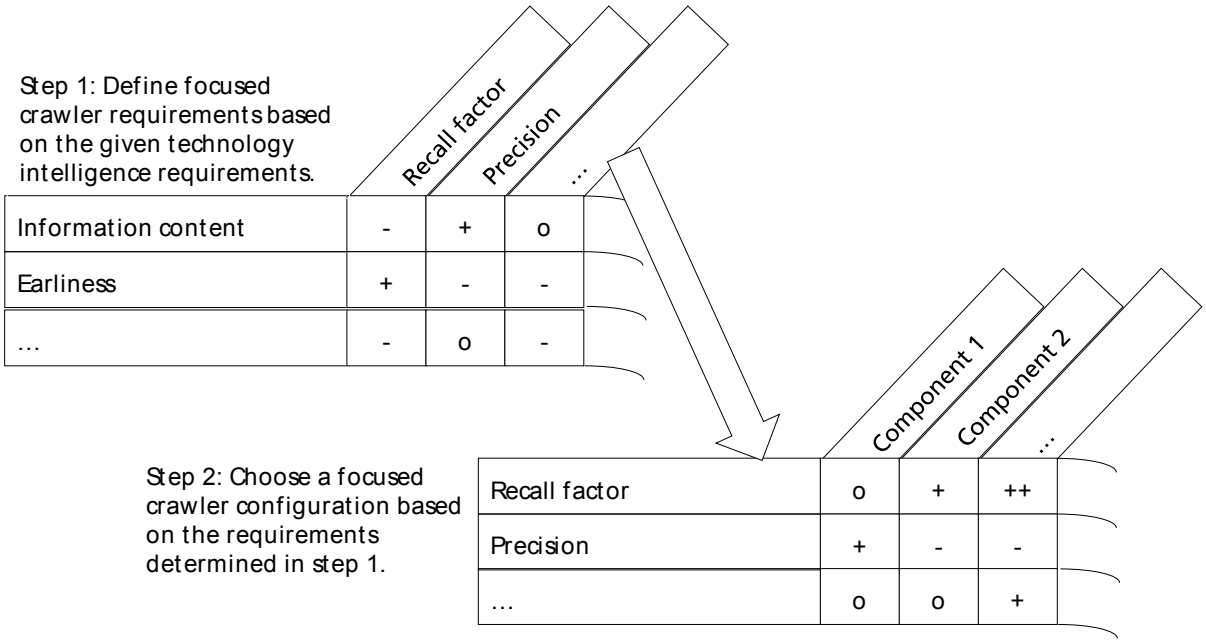

Figure 4: First concept of a focused crawler configuration matrix.

Figure 4 shows an early concept which outlines a focused crawler configuration matrix. Using such a configuration model, the requirements which are based on the current search field have to be translated into the specific task requirements towards focused crawlers (step 1). Following, the appropriate focused crawler components like algorithms and training data sets can be chosen to build a practical crawler solution (step 2).

CURRENT FINDINGS

The requirements towards technology intelligence processes and towards focused crawlers can be harmonized with each other. Focused crawlers built of different components (relevance rating algorithms, content parsing and extraction algorithms, training data, etc.) can be assessed by recall, precision and cost. After this assessment is done related to the requirements and factors of technology intelligence processes, an optimized focused crawler configuration to solve the specific task can be determined.

CONCLUSION

Because of both, the amount and the heterogeneity of the meanwhile available information, automated or computer-aided processing methods are indispensable. By a reasonable integration of existing IT solutions, first approaches to optimize research processes can be implemented. On this basis, new and continuative solutions can be developed. This demands the admission of such solutions as part of the particular research field which intends to handle the challenges of information overload. Already in this early state, the thoughts about focused crawlers in technology intelligence show that existing processes can be supported significantly.

OUTLOOK

Next, the early concept of a configuration matrix has to be advanced to a first usable draft, which allows an intuitive configuration of a required focused crawler solution. This matrix will be validated and further optimized in at least two case studies. On the one hand, it will be used in a typical technology intelligence process as mentioned in this article. On the other hand, a long-term test will be passed with a by and by optimized focused crawler running to support research activities in the Cluster of Excellence Integrative Production Technology for High-Wage Countries.

ACKNOWLEDGMENTS

This work was performed as part of the Cluster of Excellence Integrative Production Technology for High-Wage Countries, which is funded by the excellence initiative by the German federal and state governments to promote science and research at German universities.

REFERENCES

Batsakis, S., Petrakis, E. G. M., and Milios, E., (2009), Improving the Performance of Focused Web Crawlers. Data & Knowledge Engineering, 68(10), pp. 1001–1013.

Chakrabarti, S., Berg, M. van d., and Dom, B., (1999), Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks, 31, pp. 1623–1640.

International Data Corporation, (2014), The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, http://www.emc.com/leadership/digital-universe/2014iview/index.htm [5 Nov 2104].

Lichtenthaler, E., (2005), The Choice of Technology Intelligence Methods in Multinationals: Towards a Contingency Approach. International Journal of Technology Management 32(3/4), pp. 388-407.

Manning, C. D., Raghavan, P., Schütze, H., (2008), Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Micarelli, A., and Gasparetti, F., (2007), Adaptive Focused Crawling. In The Adaptive Web, P Brusilovsky, A Kobsa, W Nejdl (eds.), pp. 231-262. Berlin: Springer.

Pal, A., Tomar, D. S., Shrivastava, S. C., (2009), Effective Focused Crawling Based on Content and Link Structure Analysis. International Journal of Computer Science and Information Security (IJCSIS) 2(1).

Pant, G., and Srinivasan, P., (2005), Learning to Crawl: Comparing Classification Schemes. ACM Transactions on Information Systems 23(4), pp. 430–462. Pittsburgh: ACM.

Rawat, S., Patil, D. R., (2013), Efficient Focused Crawling based on Best First Search. 3rd International Advance Computing Conference (IACC), pp. 908-911. Ghaziabad, IEEE.

Saxler, J., (2011), Gestaltungsmodell für Netzwerke zur Technologiefrüherkennung. Aachen: Apprimus.

Schuh, G., Orilski, S., and Wellensiek, M., (2009), Efficient Technology Intelligence by Search Field Strategies. The XX. International Society for Professional Innovation Management ISPIM Conference. Wien : ISPIM.

Schuh, G., Bräkling, A., and Apfel, K., (2014), Identification of Requirements for Focused Crawlers in Technology Intelligence. Proceedings of PICMET ’14. San Jose : IEEE.

Singh, S., and Tyjagi, N., (2013), A Novel Architecture of Mercator: A Scalable, Extensible Web Crawler with Focused Web Crawler. International Journal of Computer Science and Mobile Computing (IJCSMC) 2(6), pp. 244-250.

Spath, D. (Ed.), Schimpf, S., and Lang-Koetz, C., (2010), Technologiemonitoring. Stuttgart: Fraunhofer IAO.

Srinivasan, P., Menczer, F., and Pant, G., (2005), A General Evaluation Framework for Topical Crawlers. Information Retrieval, 8(3), pp. 417–447.

US Library of Congress, (2013), Annual Report of the Librarian of Congress for the fiscal year ending September 30, 2012.

Wellensiek, M., Schuh, G., Hacker, P. A., and Saxler, J., (2011), Technologiefrüherkennung. In Technologiemanagement, G Schuh, S Klappert (eds.), pp. 89-169. Berlin: Springer.

Zhuang, Z., Wagle, R., Lee Giles, C., (2005), What’s There and What’s Not? Focused Crawling for Missing Documents in Digital Libraries. Digital Libraries, 2005. Proceedings of Joint Conference of Digital Library (JCDL ’05), pp. 301-310.