Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 151 (2019) 1231–1236

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Web Search and Data Mining (WSDM) April 29 - May 2, 2019, Leuven, Belgium

Design and Development of a Custom System of Technology Surveillance and Competitive Intelligence in SMEs

Jesus Silva[a]*, Lucelys del Carmen Vidal Pacheco[b], Kevin Parra Negrete[c], Johana Cómbita Niño[d], Omar Bonerge Pineda Lezama[e], Noel Varela[f]

[a] *Universidad Peruana de Ciencias Aplicadas, Lima 07001, Peru*
[b,c,d,f] *Universidad de la Costa (CUC), Barranquilla 080003, Colombia*
[e] *Universidad Tecnológica Centroamericana (UNITEC), Tegucigalpa 11101, Honduras*

Abstract

Making strategic decisions is a complex process that requires reliable and up-to-date information. It is therefore necessary to have tools that facilitate the information management. Technology Surveillance (TS) and Competitive Intelligence (CI) are two disciplines that seek to obtain accurate and up-to-date information. Clearly, the web is the largest and most important source of information, but their destructuring and disorganization requires tools that help to manage it. This work presents a model for TS and CI using Web Mining techniques such as ranking algorithm of web pages based on machine learning, i.e. the Advanced Cluster Vector Page Ranking (ACVPR) algorithm.

Keywords: Web mining, technology surveillance and competitive intelligence, decision making, advanced cluster vector page ranking.

1. Introduction

With the globalization of markets and the development of the digital era, the value that companies attributed to the information has been changing, becoming more strategic than ever. This means that disciplines such as Technological Surveillance (TS), and the Competitive Intelligence (CI) have become an essential basement to create

* Corresponding author. Tel.: +51920287620
E-mail address: jesussilvaUCP@gmail.com

10.1016/j.procs.2019.04.177

new products or services, to define marketing strategies, to enhance the capabilities of the organization, to improve customer service, etc. (T. Hiltbrand, 2010)[1], (Gaitán-Angulo M. et al; 2018)[2]. There are many TS methods which can be grouped into the following classes (A. Firat et al, 2008)[3]: trend analysis, expert opinion (S. Madnick and W. Woon, 2009)[4], monitoring and intelligence methods (Adamopoulos, P, 2014)[5], statistical methods, scenario modeling, and modeling and simulation methods (Ahmad, M. Wet al, 2017)[6].

CI is a discipline that is responsible for the collection, analysis, interpretation and dissemination of data about the competitive environment where companies are involved, through a systematic and ethical process, indispensable for the TS (R. Barainka, 2006)[7]. There are three stages (Xiang, B. et al, 2010)[8]: a) information collection, b) information extraction, and c) contextualization of the information. The objectives of the use of an information system of CI highlight three aspects: a) improving the competitiveness of the company, b) predicting, with a high level of confidence, the evolution of the environment, and c) providing good support for the strategic decision-making process (T. Hiltbrand, 2010)[1], (Yang, Y. et al, 2012)[9].

For its part, Web Mining (WM) is a methodology for retrieving information that allows to process and capture useful information from web pages and documents on the Internet, that can help to carry out processes of TS and CI, contributing in one of their main stages which is the search and collection of information for later analysis and treatment. Basically, the WM algorithms use data mining techniques to discover and extracting information from documents and resources available on the Web (Zhu, H. et al, 2017)[10] WM can be divided into three categories: WM of web content, WM of the structure of the website, and WM of the use of the web (Zhou, D. et al, 2018)[11] of these three categories, the most relevant to TS and CI is the first one, which covers the discovery of resources, documents categorization and clustering, and information extraction from web pages (Alam, M. and Sadaf, K., 2015)[12], (Ferretti, S. et al, 2016)[13], (I. Popa and G. Cucui, 2009)[14].

This article describes a software prototype based on WM techniques to make TS and CI in the field of SMEs. The above can be achieved by the use of the Advanced Cluster Vector Page Ranking (ACVPR) algorithm. This algorithm provides the user with a powerful meta-search tool to assist him by providing a ranking order of the web page to quickly meet custom needs, especially when the search is erroneous or incomplete.

2. System Design

The system design for the proposed metasearch tool has three sub-phases. The detailed description of each phase and a simplified block diagram are shown in Fig 1.

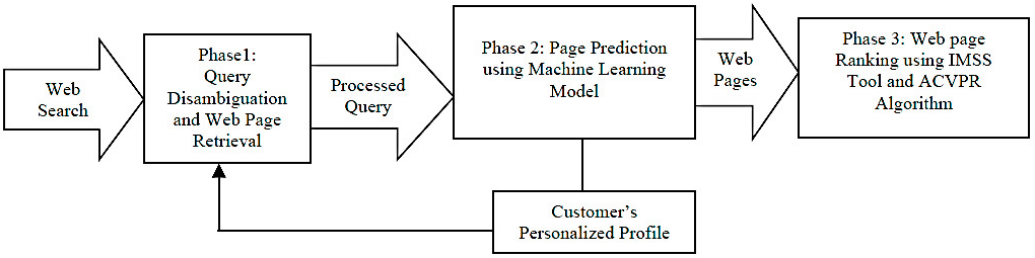

Fig. 1. System design of proposed IMSS-P tool and ACVPR (Malhotra, D. et al, 2017)[15]

The recommendation engine module can be used to build a user's profile using web technology semantics. The expanded custom query is also passed to the number of search engines in the background in the metasearch tool. To predict a user's preference for a specific web page, an automatic learning model based on logistic regression was developed. The response variable to be predicted is the feedback with respect to the relevance of the classified web link in the output of the metasearch tool. The data must be in .csv format as required by the Statistics R tool. The .csv

format file will consist of the data about the following five variables (Malhotra, D. and Rishi, O., 2016)[16], (Malhotra, D. and Rishi, O., 2017)[17], (Kamatkar S. et al; 2018[b])[18]:

- Comments represent the user relevance response for the previous web link in his browsing history and can take either of two values, Yes or No.
- The load represents the load experience of the user's web page and can take one of the two values, Good or Bad.
- The response represents the user's response time experience and can take one of two values, Good or Bad.
- Security represents the security protocol function provided by the candidate website and can take either of two values, Yes or No.
- Custom represents the use of the function, that is, the custom expansion of the function.

3. Results and Discussions

Based on the approach of TS and CI, this model is proposed to meet two objectives. The first one, generating a process to support the collection of user requirements, and the second one, generating a process of continual search for resources on the web. Both of these processes work together in a same flow, interacting constantly (Zhang, G., et al., 2012)[19]. The idea of the first process is to guide the user in the assembly of several keys that will be taken as the starting point of the search, and the second one is to provide the user with resources that are found.

The developed system has two main components, the Module for Collecting User Requirements and the WM module. Both modules interact using a Web service with which the communication process can be managed (Bouadjenek, M. et al, 2016)[20]. The user starts interacting with the Requirement Collection Module that, through a series of questions, guides the user in the construction of search keys. These keys are sent to the WM module which consults the public interfaces of four search engines: Google, Bing, Intelligo (a patent search engine) and Msmlx Excite (a metasearch engine) (Aoki, Y. et al, 2015)[21]. The first ten URLs of each are categorized and arranged in a single list according to the user´s requirements (the repeated URLs are removed from that list). From this point, the WM module begins a continuous process of inspection of links and resources on the basis of each of the URLs in the list. As the resources that are considered relevant for the user are found, they are placed in a directory along with the resource metadata. Based on the characteristics of this ongoing process (which generates a large volume of information), only fifty resources are selected and shown to the user for their evaluation.

The mining process starts when the sub-module called Web Miner (Fig. 2) receives the set of keys generated by the Requirement Collection module (Malthankar, S. and Kolte, S., 2016)[22], (Malhotra, D. and Rishi, O., 2018)[23], (Liu, Y. et al, 2017)[24], (Lis-Gutiérrez J. et al; 2018)[25]. These keys are sent to the search sub-module that is responsible for obtaining the URL provided by the public interfaces of the four search engines used (the results obtained by Google and Bing are obtained through their APIs, while Msmlx Excite and Intelligo results are obtained by simulating a manual search procedure). From the set of URLs obtained, a single filtered list is generated according to their position in the search engines and the relationship with the requirements set by the user.

For each URL, a graph of n discovered URLs is generated (Chen, C. and Zhang, C., 2014)[26], (Kamatkar S. et al; 2018[a])[27]. The edges correspond to relations between the URLs. The graph is sent to the Information Retrieval Algorithm sub-module to establish a score for each URL of the graph with the objective of obtaining a list sorted according to the relevance of content. The weighted graph is sent to the Web Scraping sub-module that is responsible for downloading the contents of the URLs that are most relevant and generate the metadata that contains information on the weighting done and the URL of the document. This information is sent to the Document Generator sub-module which creates the documents depending on the extension (.pdf, .html, .asp, .php, etc.) and link them to the

metadata to store it in a directory that will be synchronized in all locations that require this information (Owncloud Service). Then, the Web Crawler sub-module gets the following URL from the list and restarts the process.

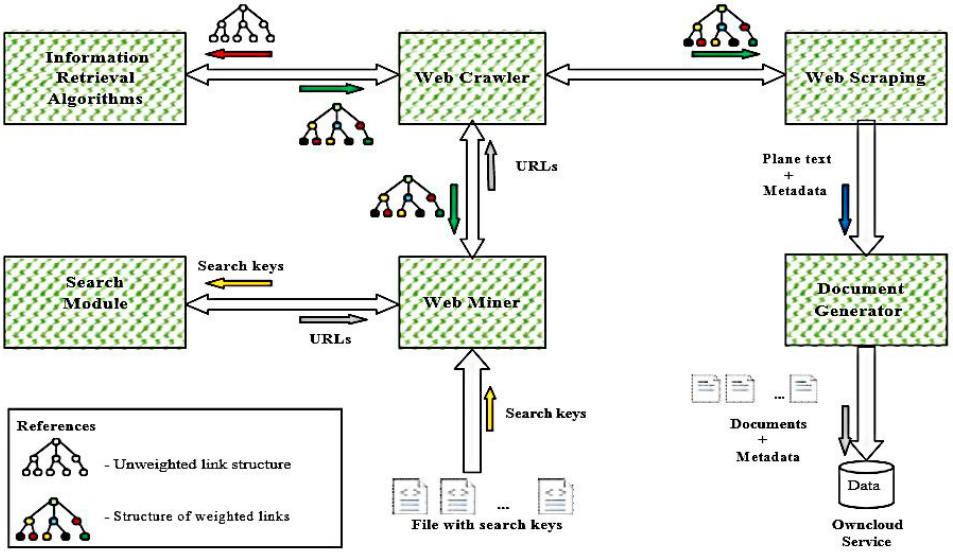

Fig. 2. Basic architecture of the application

With the objective of verifying the developed system operation, a test was performed about the information search on the productive process of the artisanal coffee. The results presented are those obtained after 12 hours from the beginning of the process. To do this, the requirement collection module was used with the following instructions: ***The production process of the artisanal coffee***, ***optimization, patents***, ***item***, ***appointment***, ***cost reduction***, ***automation*** *and* ***machinery***, excluding ***advertising***.

Table 1. Percentages of sites and resources obtained for tea exports

Type of sites	Obtained percentage	Type of resource	Obtained percentage
Commercials	64	Journal Article	15
Governmental	4	Patent	15
Academics, educational	22	Slide	10
Other	10	Commercial Information	60

The first aspect that stands out in the results of Table 1 is the high percentage of commercial sites found. Among them there is a great variety of domains that include statistical information of export of products (not only coffee) of the different countries. One of the sites on the list (theteadetective.com) appears several times with different pages, when reviewing the link it was discovered that this page has a large amount of information on the different aspects of the history, production, processing and marketing of coffee.

4. Conclusions

This research proposes the ACVPR algorithm and architecture of a meta-search system, IMSS-P for the design of a system of surveillance technology and competitive intelligence for the SMEs sector. The system has two main modules. The first one helps to guide the user in the collection of requirements for the search process generating

more comprehensive search keys than the ones used in the search engines. The second module is responsible for the Web Mining, exploring links from URLs returned by the public interfaces of the most used search engines.

One aspect of relevance of the proposed model is that it does not do a traditional query-response search, but that the results obtained are refined continuously and categorized according to user requirements. This means that, once the search process has been started by the user, the same continue running obtaining new results until the user decides to finish it.

Although the analysis carried out in this article has been quantitative due to the fact that have presented the first experiences with the system, the tests show that the system finds information according to the type of expected response. In this sense, the three tests had different objectives and the resources obtained by the system coincide with said objectives.

References

[1] T. Hiltbrand; "Learning Competitive Intelligence From a Bunch of Screwballs", Business Intelligence Journal, vol: 15, no: 4, 2010.

[2] Gaitán-Angulo M. Amelec Viloria, Jenny-Paola Lis-Gutiérrez, Dionicio Neira, Enrrique López, Ernesto Joaquín Steffens Sanabria, Claudia Patricia Fernández Castro. (2018) Influence of the Management of the Innovation in the Business Performance of the Family Business: Application to the Printing Sector in Colombia. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.

[3] A. Firat, W. Woon, and S. Madnick, "Technological Forecasting – A Review," presented at the Working Paper CISL# 2008-15, Cambridge, 2008.

[4] S. Madnick and W.L. Woon; Technology Forecasting Using Data Mining and Semantics, MIT/MIST Collaborative Research, 2009.

[5] Adamopoulos, P., 2014. On discovering non-obvious recommendations: Using unexpectedness and neighborhood selection methods in collaborative filtering systems. Proceedings of the 7th ACM international conference on Web search and data mining, ACM, 655 - 660.

[6] Ahmad, M. W., Doja, M. N., & Ahmad, T., 2017. Enumerative feature subset based ranking system for learning to rank in presence of implicit user feedback. Journal of King Saud University-Computer and Information Sciences. Elsevier

[7] R. Barainka; "Modelos de Vigilancia Tecnológica e Inteligencia Competitiva". Servico Zaintek de BAI. 2006.

[8] Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., & Li, H., 2010. Context-aware ranking in web search. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, 451-458.

[9] Yang, Y. F., Hwang, S. L., & Schenkman, B.,2012. An improved Web search engine for visually impaired users. Universal Access in the Information Society, 11(2), 113-124.

[10] Zhu, H., Ou, C. X., Van den Heuvel, W. J. A. M., & Liu, H.,2017. Privacy calculus and its utility for personalization services in e- commerce: An analysis of consumer decision-making. Information & Management, Elsevier, 54(4), 427-437.

[11] Zhou, D., Zhao, W., Wu, X., Lawless, S., & Liu, J., 2018. An iterative method for personalized results adaptation in cross-language search. Information Sciences, Elsevier, 430, 200-215.

[12] Alam, M. and Sadaf, K., 2015. Labeling of Web Search Result Clusters using Heuristic Search and Frequent Itemset. Procedia Computer Science, Elsevier,216-222.

[13] Ferretti, S., Mirri, S., Prandi, C., & Salomoni, P., 2016. Automatic web content personalization through reinforcement learning. Journal of Systems and Software, Elsevier, 121, 157-169.

[14] I. Popa Anica and G. Cucui, "A Framework for Enhancing Competitive Intelligence Capabilities using Decision Support System based on Web Mining Techniques", Int. J. of Computers, Communications & Control, vol. 4, no. 4, pp. 326-334, 2009.

[15] Malhotra, D., Malhotra, M. and Rishi, O.P., 2017.An Innovative Approach of Web Page Ranking Using Hadoop- and Map Reduce- Based Cloud Framework. Proceedings of Advances in Intelligent Systems and Computing, Vol.654, CSI, Springer, 421-427.

[16] Malhotra, D. and Rishi, O.P., 2016. IMSS-E: An Intelligent Approach to Design of Adaptive Meta Search System for E-Commerce Website Ranking. Proceedings of the International Conference on Advances in Information Communication Technology & Computing, ACM, doi>10.1145/2979779.2979782.

[17] Malhotra, D. and Rishi, O.P., 2017. IMSS: A Novel Approach to Design of Adaptive Search System Using Second Generation Big data Analytics. Proceedings of International Conference on Communication and Networks, Springer, 189-196.

[18] Kamatkar S.J., Kamble A., Viloria A., Hernández-Fernandez L., Cali E.G. (2018)b. Database Performance Tuning and Query Optimization. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[19] Zhang, G., Li, C. and Xing, C., 2012. A Semantic++ Social Search Engine Framework in the Cloud. In Semantics, Knowledge and Grids (SKG), 2012 Eighth International Conference, IEEE, 270-278

[20] Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., & Vakali, A., 2016. Persador: personalized social document representation for improving web search. Information Sciences, Elsevier, 369, 614-633.

[21] Aoki, Y., Koshijima, R. and Toyama, M., 2015. Automatic Determination of Hyperlink Destination in Web Index. In Proceedings of the 19th International Database Engineering & Applications Symposium, ACM, 206-207.

[22] Malthankar, S. V., & Kolte, S., 2016. Client Side Privacy Protection Using Personalized Web Search. Procedia Computer Science, Elsevier, 79, 1029-1035.

[23] Malhotra, D., & Rishi, O. P.,, 2018. An intelligent approach to design of E-Commerce metasearch and ranking system using next- generation big data analytics. Journal of King Saud University-Computer and Information Sciences, Elsevier

[24] Liu, Y., Bi, J.W. and Fan, Z.P., 2017. Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. Information Fusion, 36, 149-161.

[25] Lis-Gutiérrez JP., Gaitán-Angulo M., Lis-Gutiérrez M., Viloria A., Cubillos J., Rodríguez-Garnica PA. (2018) Electronic and Traditional Savings Accounts in Colombia: A Spatial Agglomeration Model. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[26] Chen, C. P., & Zhang, C. Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, Elsevier, 275, 314-347

[27] Kamatkar S.J., Tayade A., Viloria A., Hernández-Chacín A. (2018)a. Application of Classification Technique of Data Mining for Employee Management System. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.