DOI: 10.20396/revpibic.

XXVII Congresso de Iniciação Científica Unicamp

16 a 18 de outubro de 2019 - Campinas | Brasil

Web crawler to search for weak signals using specific prioritized sources

Ana Estela Antunes da Silva, Pedro Ivo Garcia Nunes, Pedro Artico Rodrigues*

Abstract

Competitiveness among business organizations requires the development of strategies to anticipate threats and opportunities. Competitive intelligence considers that it is possible to monitor these threats and opportunities using weak signals (WS). WS are initial information on emerging and unknown phenomena whose analysis can provide strategic knowledge to companies. Some studies suggest that Web-based technologies are suitable for WS monitoring. This project presents a Web crawler specifically dedicated to the search of WS.

Key words:

competitive intelligence, Web crawling, weak signals

Introduction

In a competitive world, organizations continually need to improve strategic planning to survive in the business market. In this context, the exploitation of opportunities and the escape of threats depends on the organizational environment monitoring (Almeida & Hirata, 2016). This monitoring can be done through techniques of competitive intelligence (CI). One of these methods involves the capturing of WS (Ansoff, 1980).

The process of WS capturing requires the development of methods to detect them (Ilmola-Sheppard & Kuusi, 2013). Web crawling is an automatic option to WS monitoring and capturing. However, conventional crawlers are not dedicated to the search for WS.

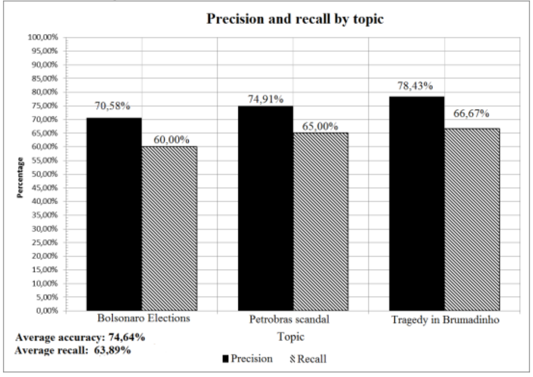
Methodology

The methods used for the project were: (i) focused Web crawler to search for WS, (ii) regular expression to clean the returned content, (iii) preprocessing techniques to obtain sentences, (iv) languages Python and PHP for implementation. The tool contains four modules that allow: management of search sources, three types of search, content display, besides selection and export of sentences.

Results and Discussion

The Web crawler has been evaluated through tests considering the metrics precision and recall. The tests carried out involved: comparison between manual selection of pages and search of pages using the Web crawler by the calculation of precision and recall. Precision and recall were calculated considering the three topics showed in the graph of Figure 1. The variation of these values is low, demonstrating that the Web crawler performed similarly when searching for WS on each of the topics.

Figure 1 – Precision and recall



Source: the author

The last step of evaluation was to verify if the average precision and recall values were consistent to the related work. Some works that proposed and evaluated focused crawlers (without any of them intending to treat WS) were selected. Table 1 shows this comparison.

Table 1 – Comparison with related work

Related work	Average precision	Average recall
Dong et.al. (2004)	77.26%	58.75%
Stamatakis et.al. (2003)	45.20%	92.10%
This work	**74.64%**	**63.89%**

Source: the author

The results of the crawler are consistent with the average precision and recall of related work considered in the comparison. It should be noted that the comparison does not take into account the fact that the related works were not specifically concerned about WS. In addition, the tool proposed by this work is more comprehensive insofar it offers functionalities besides the Web crawling.

Conclusions

The motivation for this work concerns the lack of Web crawling tools concentrated on the search for WS. In this context, this work proposed and evaluated a focused Web crawler which can be considered innovative because this tool looks specifically for WS, meeting the requirements of related works that use this type of information to detect threats and business opportunities.

Acknowledgement

The authors would like to express thanks to FAPESP for the financial support (process number: 2018/10229-9).

1. Almeida, F., Hirata, P. (2016). Entendendo e implantando um sistema de inteligência competitiva. REGE - Revista de Gestão, 111-122.
2. Ansoff, I. (1980). Strategic Issue Management. Strategic Management Journal, 1, 131-148.
3. Dong, P., et. al. (2004). Quantitative evaluation of recall and precision of CAT Crawler, a search engine specialized on retrieval of Critically Appraised Topics. BMC MedInform Decis Mak.
4. Ilmola-Sheppard, L., Kuusi, O. (2013). Information filters as one of the means of managing strategic fit in a complex environment (Vol. 15). Foresight.
5. Pal, A., et. al. (2009). Effective Focused Crawling Based on Content and Link Structure Analysis. Department of Computer Science Engineering, India.

BY-NC-ND