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 advantages. 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 shown in the graph of Figure 1.

![Figure 1 – Precision and recall](image)

Source: the author

The variation of these values is low, demonstrating that the Web crawler performed similarly when searching for WS on each of the topics.

The last step of evaluation was to verify if the average precision and recall values were consistent with 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>
<thead>
<tr>
<th>Related work</th>
<th>Average precision</th>
<th>Average recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dong et.al. (2004)</td>
<td>77.26%</td>
<td>58.75%</td>
</tr>
<tr>
<td>Stamatakis et.al. (2003)</td>
<td>45.20%</td>
<td>92.10%</td>
</tr>
<tr>
<td>This work</td>
<td>74.64%</td>
<td>63.89%</td>
</tr>
</tbody>
</table>

Source: the author

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).