DEVELOPMENT OF A FOCUSED WEB PAGE CRAWLER BASED ON GENRE AND CONTENT

Marcelo Trajano Alves Júnior, Marcos Felipe Pontes Rezende and Guilherme Tavares de Assis

Department of Computing - Federal University of Ouro Preto
Ouro Preto - MG, Brazil

ABSTRACT

Focused crawlers are generally used to crawl pages that satisfy some particular property and that are relevant to a specific topic of interest and are important for a wide variety of applications. For particular situations, a focused crawling approach was proposed and developed where the topic of interest can be expressed by terms that describe the genre and content of the desired web pages. In order to improve the efficiency and effectiveness of such an original genre-aware approach to focused crawling, the following improvements have been proposed, developed and validated: relevant page location policy based on Link Context, semi-automatic seed page determination, automatic similarity threshold definition and automatic refinement of genre and content term sets. In this context, this work proposes to develop a complete and functional version of a crawler, called Yucca, following the original genre-aware approach to focused crawling and the improvements already developed and validated, so that it can be used by different users in a simple and robust way. To validate Yucca, experiments were performed involving the crawling of web pages referring to three distinct and current topics of interest. In general, Yucca presented itself as an effective focused crawler, since the levels of precision achieved by the crawling processes carried out were quite satisfactory, reaching more than 80% on average when considering 10 pages returned as relevant by the crawler.

KEYWORDS

Focused Web Crawler, Focused Crawling Processes, Genre Terms, Content Terms

1. INTRODUCTION

Currently, according to Ahlgren (2021), there are more than 1.83 billion websites on the internet and this number grows exponentially each year; with this, it becomes necessary to create new Information Retrieval techniques, in order to facilitate the crawl Web pages and, consequently, the search for information desired by users. For this, as seen in Bhatt et al. (2015), search engines are basic tools to search for something of interest on the internet from repositories that are generated by traditional Web crawlers: a traditional Web crawler serves to crawl Web pages starting with seed pages and following the links contained in it, thus visiting other pages until it has covered a sufficient number of pages or reached a certain objective.

However, according to Costa et al. (2017), general purpose search engines do not solve well the problem of locating web pages referring to a specific topic, as the page collections generated by them are quite voluminous and, generally, user queries are short involving little information. In this context, focused crawlers (Jiang et al., 2013) serve to generate smaller and restricted page collections, as they have the larger purpose of crawling pages that are, in the best possible way, relevant to a specific topic or interest of the user, from a more detailed specification of what one wants to crawl.

Thus, aiming to perform effective and efficient processes of focused crawlers, an approach was proposed and developed (Assis et al., 2009) aimed at meeting specific situations. In general, such an approach consists of considering the evidence of genre (the type or style of text in specific documents) and content (the subject or theme you want to crawl) present on a given page and establish a degree of similarity between such evidence and the specific topic of interest. Therefore, this work had, as its main objective, to establish a framework that allows the construction of effective, efficient and scalable focused crawlers, without the need for a priori training or any type of pre-processing. Specifically, the proposed focused crawler approach is useful in situations where a topic of interest can be expressed through two distinct sets of terms: the first
describing genre aspects of the desired pages and the second referring to the subject or content described on those pages. Through experiments performed, such an approach to focused crawler based on genre presented satisfactory levels of precision, recall and F1: 85% to 100% for all topics of interest considered.

However, there is no crawler, that is, a functional tool itself, which performs focused crawling processes, following the original approach mentioned, and also includes the improvements already applied and duly validated in the approach (Mangaravite et al., 2012, Mangaravite et al., 2014, Siqueira et al., 2016, Costa et al., 2017, Assis and Souza, 2018). Thus, this work proposes to develop and validate a complete and functional version of the focused crawler based on genre and content, called Yucca, considering the original approach and the integration of components related to the already validated improvements of the approach, so that it can be used by different users in a simple and robust way. Then, the main contributions of this work are: (a) proposal of Yucca, a focused web page crawler based on genre and content of interest; (b) improving the effectiveness (determination of relevant pages) and efficiency (faster locating of relevant pages) of the crawling processes performed by Yucca compared to the original defined approach; (c) definition of functional characteristics related to the crawling processes that users wish to perform, through the use of a friendly interface proposed for Yucca; (d) analysis of results obtained through real Yucca validation experiments, involving specific topics relevant to the current moment.

The remainder of this work is organized as follows. In Section 2, related works are presented. In Section 3, the focused crawler proposed in this work, involving its functioning architecture, characteristics and layout, is described. In Section 4, the practical experiments performed are presented and the results obtained are analyzed. Finally, in Section 5, conclusions and perspectives for future work are presented.

2. RELATED WORK

As already mentioned, this work aims to develop and validate the first complete and functional version of Yucca: a focused web crawler based on genre and content. Thus, as related works, the original genre-aware approach and its proposed and developed improvements (see Subsection 2.1) and examples of current focused crawlers guided by heuristics (see Subsection 2.2) are presented.

2.1 Original Approach to Genre-Aware Focused Crawling

The original approach to genre-aware focused crawling (Assis et al., 2009) establishes a framework that allows the construction of effective, efficient and scalable focused crawlers, which take into consideration the genre and content of the desired pages. Figure 1 shows the architecture of the original approach to genre-aware focused crawling.

According to Siqueira et al. (2016), as you can see in Figure 1, firstly (step 01), the priority queue called Frontier is initialized with the URLs of the seed pages (a set of pages from which to start the crawling), setting the URL scores to 1. For each URL in Frontier (step 02), the corresponding page is visited (step 04) and its content analyzed (steps 05 to 09): each page is represented as a n-dimensional vector based on its terms (vector model) and the cosine distance is used to measure the similarity between the current page and the set of terms that represent the pages of interest. This measure is calculated separately to each set of terms (steps 05, 06 and 08), generating a specific similarity score between the current page and the sets of terms that represent, respectively, the genre, the content and the URL string of the desired pages. Each URL string term is related to the page genre or to the desired content. Then, these scores are combined into a final single one (steps 07 and 09), considering different weights for the sets of genre terms and content terms, and compared with a given threshold defined by an expert. If this final score is greater or equal to this threshold (step 10), the visited page is included in the set of relevant pages. Next, if the current page is considered relevant, the scores of URLs in Frontier that correspond to the sibling pages of the current page are changed (step 11). Finally, the previously extracted links from the current page are inserted into the Frontier (step 12) having their scores set to 0.

As already mentioned, improvements were proposed, developed and validated, in order to improve the original approach described in Figure 1. As a first improvement, the use of Link Context was proposed in (Mangaravite et al., 2012), which aims to use text anchor, link title, and URL to improve the process of determining the visit priority scores that define the ordering of unvisited URLs found in the crawler's
Frontier. In general, to compute such scores, we also used the cosine distance between the terms of genre and content, input parameters of the original approach, and the texts generated by using the Link Context. The application of such a technique resulted in the improvement of the crawler's visit policy, generating an increase of up to 100% of efficiency in the original genre-aware approach.

![Architecture of the original approach to genre-aware focused crawling (Assis et al., 2009)](image)

As a second improvement, Mangaravite et al. (2014) proposed a strategy for semi-automatic generation of seed pages, related to a certain topic of interest, so that the relevant pages to the desired topic are more quickly located by the crawler. The proposed strategy consists of using the specified terms of genre and content in a search engine, more specifically Google, to generate the seed pages. According to the experiments performed, the UnionFirst heuristic established for semi-automatic generation of seed pages, which uses only the first genre and content term in the query sent to the search engine, resulted in an improvement in efficiency in the original approach of up to 53%.

As a third improvement, in the work developed by Siqueira et al. (2016), three strategies were developed to automatically determine the similarity threshold used in focused crawling processes of the original approach. For each strategy developed, focused crawling processes were performed involving three distinct topics of interest. Through the results obtained, it was observed that the crawling processes, related to the strategy based on a K-Means grouping method (partitioning method), were the ones that presented the best effectiveness values, reaching very close F1 levels (difference of only 5.4%) from those obtained when the similarity thresholds were defined by specialists of the topics of interest considered.

And finally, as a fourth improvement, aiming to improve the sets of terms of genre and content, provided as input data, two strategies, for improving such sets based on association matrix and natural language processing, were proposed by Costa et al. (2017). Through the analysis of the results of the experiments described, it was possible to see that the strategy based on a matrix of association of terms, using the established metric Shortest Distance (calculation of the similarity sij, between two terms ti and tj, by the normalized sum of the smallest distances between these terms, considering all pages that have these terms) was the one with the best results, promoting an increase in the F1 metric of 6.29% when compared to the F1 value obtained by the crawling process, for the same specific topic, whose terms of genre and content have not been expanded.

### 2.2 Focused Crawlers Guided by Heuristics

In Lee et al. (2019), a genre-aware focused crawler was proposed and developed (in this case, genre refers only to academic texts), called SlideCrawler, aiming to crawl slide files with academic content, through Google as a crawling tool to manage queries and perform the desired downloads. The proposed crawler has: (a) a query generator to specify the desired slide format and the university to be consulted; (b) a URL
extractor that is responsible for extracting URLs from slides and removing possible duplicates; and (c) a download manager that downloads the files pointed to by the extractor. In the experiments performed, SlideCrawler was able to download more than 850,000 academic slide files with diverse content. Comparing with another crawler called Apache Nutch (open source web crawling tool), SlideCrawler was able to crawl 3.7 times more slide files. However, despite the crawler being based on genre, it is limited to only a file format and a specific site, unlike the crawler proposed in this work, which aims to crawl the largest amount of pages in a given specified topic.

Not considering genre, Chen et al. (2012) use a recognition algorithm based on link analysis to obtain the most relevant pages to the desired topic of interest. This algorithm follows two premises: (a) if page A has a link to page B, then page B is a recommendation for page A; and (b) if there are links that connect pages A and B, then both pages can belong to a common theme. Based on this, Chen et al. (2012) deduced two more premises, namely: (a) if pages A and B point to the same pages, then these two pages are considered relevant, that is, the more links two pages match, the greater the degree of relevance between them; and (b) if a page has many links pointing to the same topic, it means that this page has a high chance of being relevant to the topic as well. Thus, to consider that a particular page A visited by the proposed focused crawler is relevant to the specified theme, it is necessary that the ratio between the number of links that are on such page A and the number of links that lead to it is greater than a predefined threshold. Considering medical themes in their experiments, this approach obtained a level of precision higher than 93% and recall higher than 83%, considering similarity thresholds equal to 0.5, 0.6, 0.7, 0.8 and 0.9; however, the approach does not use the semi-automatic generation of similarity threshold and seed pages, unlike the crawler proposed in this work, which has such functionalities in order to improve the effectiveness and efficiency, without the intervention of users regarding the provision of similarity threshold and seed pages, from focused crawling processes.

3. PROPOSAL AND DEVELOPMENT OF YUCCA

From the original approach to genre-aware focused crawling (see Figure 1) and its presented improvements, a functional and complete crawler, called Yucca, to focused crawling based on genre and content was proposed and developed. Figure 2 shows the functioning architecture of Yucca.

According to Figure 2, aiming at a particular crawling process to be performed for a specific topic of interest, the terms of genre (Step 01) and content (Step 02) are initially specified, these being the user's tasks. Then, in Step 03, the seed pages are semiautomatically generated using the terms of genre and content specified; such seed pages initialize the list of unvisited URLs, present in Frontier, with Yucca's visit priority score equal to 1. Considering the generated seed pages, Step 04 generates the association matrix for definition (Step 05) of terms expanded of the original terms. Continuing, in Step 06, the similarity threshold is automatically specified and determined using the terms specified by the user in Step 01 and 02 and the expanded terms in Step 05. Starting the crawling process itself, while there are unvisited URLs in the Frontier (Step 7), the one with the highest visit score is unqueued from Frontier (Step 08) and the corresponding page is visited by Yucca (Step 09); this visit consists of analyzing its relevance, through a set of similarity calculation heuristics, regarding the specific topic of interest. Thus, in Steps 10 and 11, the cosine distances between the visited page and the original and expanded terms of genre and content are calculated, respectively, combining and generating, in Step 12, the similarity of genre and content. Then, in Step 13, the cosine distance between the original and expanded terms of genre and content and the URL of the visited page is calculated, combining it, in Step 14, with the calculated similarity of genre and content (Step 12), thus generating the final similarity of the page visited in relation to the specific topic of interest. If such final similarity is greater than the automatically generated similarity threshold (Step 15), the visited page is considered relevant and, thus, it is stored in the repository of relevant pages to the specific topic of interest; in addition (Step 16), according to the queuing policy defined for Frontier, the visit score of URLs not yet visited, corresponding to the sibling pages of the visited page, is changed to the value of the calculated final similarity. Finally (Step 17), not linked to the execution of heuristics to calculate similarity, the URLs present in the visited page are added to Frontier with visit scores defined by the similarity between the terms of genre and content and the link contexts (Mangaravite et al., 2012) of the URLs in question.
4. PRACTICAL EXPERIMENTATION

In this section, the Yucca evaluation experiments are presented and analyzed, following the architecture proposed in Figure 2. Subsection 4.1 describes the performed experiments and Subsection 4.2 presents and evaluates the results obtained through the performed experiments.

4.1 Experimental Setup

In order to evaluate the first functional version of Yucca, crawling processes were performed considering 3 current and distinct topics of interest, namely: (1) articles related to symptoms caused by Covid-19; (2) articles related to structural racism; (3) and articles related to global warming. Furthermore, in order to verify the importance of the content and genre terms used in the crawling processes, 3 different weight combinations were considered for the genre and content terms: Genre 0.3 and Content 0.7; Genre 0.4 and Content 0.6; and Genre 0.6 and Content 0.4. Due to the fact that the three defined topics have the same genre (articles), the same set of genre terms was specified for the topics, namely: article, introduction, conclusion, theoretical framework, abstract and result. Regarding content terms, different sets of terms were specified for each topic of interest, namely: (topic 1) covid-19, symptoms, signs and effects; (topic 2) structural racism, prejudice and racial discrimination; (topic 3) global warming, climate, climate change, ozone layer, greenhouse effect, temperature and environment. For all crawling processes performed, the following common characteristics were specified: maximum number of pages visited: 5000; maximum number of pages returned by Yucca, as relevant, to calculate precision: 60; and weight of a page's URL and genre/content combination (used in Step 14 of Figure 2): 0.5.

Furthermore, in order to analyze the pages returned as relevant by Yucca, throughout the execution of each crawling process, a log was stored containing the following information about each visited Web page: visited page identifier, automatically assigned by the crawler; URL of the visited page; HTML code of the visited page; and calculated similarity value between the page visited and the original and expanded genre and content terms for the topic of interest.

Finally, to evaluate the experiments performed, the precision metric was used. According to Brownlee (2020) and considering the context of this work, precision is a metric that establishes the fraction of pages really relevant to the desired topic of interest, which were returned by the focused crawler, in relation to all pages returned by it.
4.2 Experimental Results

Considering all the crawling processes performed, for each topic of interest defined, Table 1 presents the test case (weights associated with the terms of genre and content), the similarity threshold reached, the number of pages visited and the number of pages returned and therefore considered relevant by Yucca. Note that such values are presented for each test case performed for the same topic of interest, being: (1) "articles related to symptoms caused by Covid-19", (2) "articles related to structural racism" and (3) "articles related to global warming", varying the weight of the genre and content terms.

Table 1. Results of the test cases performed

<table>
<thead>
<tr>
<th>Topic</th>
<th>Test Case</th>
<th>Similarity threshold</th>
<th>Number of visited pages</th>
<th>Number of retrieved pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Genre: 0.3/ Content: 0.7</td>
<td>0.3251</td>
<td>2838</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Genre: 0.4/ Content: 0.6</td>
<td>0.2663</td>
<td>2876</td>
<td>460</td>
</tr>
<tr>
<td></td>
<td>Genre: 0.6/ Content: 0.4</td>
<td>0.3365</td>
<td>3007</td>
<td>63</td>
</tr>
<tr>
<td>(2)</td>
<td>Genre: 0.3/ Content: 0.7</td>
<td>0.2189</td>
<td>3876</td>
<td>3037</td>
</tr>
<tr>
<td></td>
<td>Genre: 0.4/ Content: 0.6</td>
<td>0.3796</td>
<td>3845</td>
<td>1198</td>
</tr>
<tr>
<td></td>
<td>Genre: 0.6/ Content: 0.4</td>
<td>0.4503</td>
<td>3916</td>
<td>177</td>
</tr>
<tr>
<td>(3)</td>
<td>Genre: 0.3/ Content: 0.7</td>
<td>0.1335</td>
<td>3646</td>
<td>2082</td>
</tr>
<tr>
<td></td>
<td>Genre: 0.4/ Content: 0.6</td>
<td>0.3716</td>
<td>3577</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>Genre: 0.6/ Content: 0.4</td>
<td>0.3716</td>
<td>3564</td>
<td>1750</td>
</tr>
</tbody>
</table>

Figures 3, 4 and 5 show, for each topic of interest, the levels of precision obtained considering different amounts of pages retrieved by Yucca, in descending order of similarity to the desired topic: 5 to 60 pages returned, from 5 out of 5. For instance, considering an arbitrary topic and a number of pages retrieved k, the graph points the fraction of the retrieved pages by Yucca that is relevant indeed.

As can be seen in Figure 3, related to the crawling processes associated with the topic of interest "articles related to symptoms caused by Covid-19", test case 1, associated with weights of 0.3 for genre and 0.7 for content, obtained a precision higher than the other tests, maintaining an average level of 83% when considering the 60 pages returned. However, when considering only the first 10 pages returned with greater similarity by Yucca, a common case in a search engine, test 3, associated with weights of 0.6 for genre and 0.4 for content, presents an average precision higher than the others, achieving 90% of precision. Moreover, the test case 3, associated with weights of 0.6 for genre and 0.4 for content, produces 100% of precision with 5 pages returned. However, the results of both test cases 2 and 3 get worse considering more pages.

Considering the topic of interest "articles related to structural racism", as can be seen in Figure 4, the precision curves for each test case were very similar; however, test case 1, associated with weights of 0.3 for genre and 0.7 for content, was slightly superior to the other tests, maintaining an average level of precision of 84% when considering the 60 pages returned. All tests showed satisfactory levels of precision, regardless of the number of pages returned, which can be seen in the graph with its lines very close to each other.
Regarding the topic of interest "articles related to global warming", as seen in Figure 5, test 3, associated with weights 0.6 for genre and 0.4 for content, was slightly superior to the others, maintaining an average level 82% precision when considering the 60 pages returned. When considering the first 10 pages returned with greater similarity by Yucca, tests 2 and 3 have similar averages of precision close 85%. Furthermore, with a high level of pages returned, both test cases presented a high precision close to 75%.

Figure 3. Precision levels – Covid-19 topic
Figure 4. Precision levels - Structural Racism topic
Figure 5. Precision levels - Global Warming topic

Figure 6 shows the best curve obtained by the three topics of interest using the most accurate test case for each one. Comparatively, it is observed that the generated precision curves remained very close and with good levels of precision, thus demonstrating accurate and satisfactory results for the topics considered. It is noteworthy, in this case, the crawling process related to the topic of interest "articles related to symptoms caused by Covid-19", since, despite having obtained the worst levels of precision for the first 10 pages returned, it presented about 90% precision for the first 40 pages returned as relevant by Yucca.

Figure 6. Precision comparison between the best results of the topics
5. CONCLUSION

As presented, this work proposes to develop a complete and functional version of a focused crawler based on genre and content, called Yucca, considering the original approach proposed in Assis et al. (2009) and the improvements made by Mangaravite et al. (2012, 2014), Siqueira et al. (2016) and Costa et al. (2017).

Seeking to evaluate the Yucca, as seen, experiments were performed considering 3 distinct topics of interest and, in all topics, the efficacy results were very satisfactory, with similar levels of precision. In particular, it was possible to observe that, depending on the weights of the terms of genre and content, the levels of precision can be different, although, regardless of such weights, the levels of precision were above 80% for up to 10 pages returned as relevant by Yucca in the three topics. This is an excellent result since, when analyzing documents linked to a specific topic, users generally check the first documents returned.

As future works, we intend to (1) propose, develop and integrate to Yucca a component for semi-automatic determination of terms of genre and content, linked to a specific topic of interest, necessary for carrying out a crawling process; (2) perform new Yucca validation experiments using, including, other metrics such as recall and F1; and (3) conduct user experience studies regarding the use of Yucca, in order to analyze its usability.

ACKNOWLEDGEMENT

This research was partially funded by research grants from PROPP/UFOP. Furthermore, it was carried out on the GAID/UFOP Laboratory.

REFERENCES


