IADIS International Journal on WWW/Internet
Vol. 15, No. 1, pp. 15-30
ISSN: 1645-7641

STRATEGIES FOR AUTOMATIC DETERMINATION OF SIMILARITY THRESHOLD FOR GENRE-AWARE FOCUSED CRAWLING PROCESSES

Gustavo Oliveira de Siqueira[1], Guilherme Tavares de Assis2, Anderson Almeida Ferreira[2], Amanda Sávio Nascimento e Silva[2], Vítor Mangaravite[1] and Flávio Luis Cardeal Pádua[3]
[1]*Department of Computing – Universidade Federal de Minas Gerais – Belo Horizonte, MG, Brazil*
[2]*Department of Computing – Universidade Federal de Ouro Preto – Ouro Preto, MG, Brazil*
[3]*Department of Computing – Centro Federal de Educação Tecnológica de Minas Gerais – Belo Horizonte, MG, Brazil*

ABSTRACT

The great popularity and, specially, the fast Web growth have led to the proposal and analysis of new techniques for helping users to locate effectively the needed information in a satisfactory time, without much difficulty. Traditional crawlers are not capable to identify relevant sub-spaces on Web related to a specific theme; however, focused crawlers are capable to solve, effectively and efficiently, the mentioned problem. Usually, a focused crawler process requires a specific value, called similarity threshold value, for determining whether a crawled Web page is relevant or not according to a topic of interest; such value is distinct for each specific topic. In order to determine automatically such a value for focused crawlers related to a genre-aware approach, we propose three strategies in this work. Our experimental evaluation achieved, as the best result, 100% of precision and 98% of F1, considering a specific crawling process for which it was determined automatically a similarity threshold value: a great result compared with the baseline.

KEYWORDS

Similarity threshold, web crawling, focused crawling

1. INTRODUCTION

Web's intense growth due its great popularity have led to new techniques proposal and analysis in order to help users locating the needed information effectively in a satisfactory time, without greater difficulties. Information access on Web is basically made through search engines that exploit the Web's graph structure in order to locate relevant pages to a specific search (Menczer et al. 2004). To allow this, a typical search engine uses a specific program, named crawler, which crawls the largest possible number of Web pages, generating an indexed pages collection. In this case, a traditional crawler crawls Web pages starting by seed-pages set and following the links on it, visiting other pages. Such a process continues with new pages, whose new links are followed until get a sufficient number of pages or reach some specific goal. Another program, named indexer, reads the collected pages and create an index based on the words on the pages. Each search engine uses its own algorithm to create its index aiming to return relevant results for user's searches.

The challenge of identifying specific and relevant Web sub-spaces, according to a theme, is typically referred to intelligent construction in a crawling strategy (Pant and Srinivasan, 2005). This intelligence is usually achieved by the use of appropriated heuristics, which guide the crawling process and aim at determining whether a page is relevant or not to a specific topic of interest. In this case, the crawlers are named focused or topical crawlers. Many focused crawler strategies (Almpanidis et al., 2007; Chakrabarti et al., 1999; Johnson et al. 2003; Pant & Srinivasan, 2005, 2006) make use of text classifiers to determine such relevance with additional cost of having to train the classifiers. Furthermore, due to generality of the situations where these strategies are applied, the crawlers reach low recall and precision levels: generally between 40% and 70%.

On focused crawler context, the work presented in Assis et al. (2009) (see Subsection 2.1) had, as main goal, propose a framework that allows the construction of effective, efficient and scalable focused crawlers, requiring no training or any kind of preprocessing. Specifically, the proposed approach to focused crawling in Assis et al. (2009) is useful in situations where the topic of interest can be expressed by two distinct set of terms: the first set describing genre aspects of the desired page's genre[1] and the second related to the subject or content of these pages. Therefore, following this approach, a specialist must specify sets of content and genre terms to represent the topic of interest and also a similarity threshold value that defines whether a visited page by the crawler is relevant or not to a specific topic. Depending on the similarity threshold value, the effectiveness of a crawling process may not be satisfactory. In order to validate the proposed approach, the authors defined empirically a similarity threshold value for each desired topic. Thus, to become the approach less user dependent, it is necessary to determine automatically the similarity threshold value to be used in a crawling process, according to a specific topic of interest, in order to ensure the effectiveness of such crawling process.

The solution for the mentioned problem is not trivial, once the similarity threshold value for a focused crawling process depends directly on the topic of interest. To achieve this goal, we propose different strategies to automatically determine the similarity threshold, based on statistical data summarizing techniques, traditional clustering methods and the silhouette coefficient metric (Pant et al., 2004).

The rest of this paper is organized as follows. Section 2 addresses related work. Section 3 presents the proposed strategies for automatic determination of similarity threshold values to be used in focused crawling processes. Section 4 describes the experimental evaluation on the defined strategies and the results obtained. Finally, Section 5 concludes the paper and gives some directions for future work.

2. RELATED WORK

Our proposed strategies, in this work, aims to determine automatically similarity threshold values to be used on focused crawling processes following a genre-aware approach. Therefore, works related to this paper are divided in two subsections: the genre-aware approach to focused crawling and automatic threshold estimation for data matching applications.

2.1 Genre-aware Approach to focused Crawling

The genre-aware approach to focused crawling presented in (Assis et al. 2009; Assis et al. 2007), used as base for this work, considers a set of heuristics to guide a crawler, in such way that it allows the separate analysis of the genre and the content of a Web page. The set of heuristics has been designed with two main objectives: improving the level of F1 of the crawling process (effectiveness) and speeding up the crawling of relevant pages (efficiency).

The Figure 1 illustrates, according to (Assis et al. 2009; Assis et al. 2007), the functioning model of the genre-aware approach to focused crawling. As we can see, firstly (step 01), a 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). It has used the cosine measure (Baeza-Yates & Ribeiro-Neto, 1999) to determine 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) 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 pages siblings of the current page are changed to final score (step 11). Finally, the links previously extracted from the current page are inserted into Frontier (step 12) having their scores set to 0.

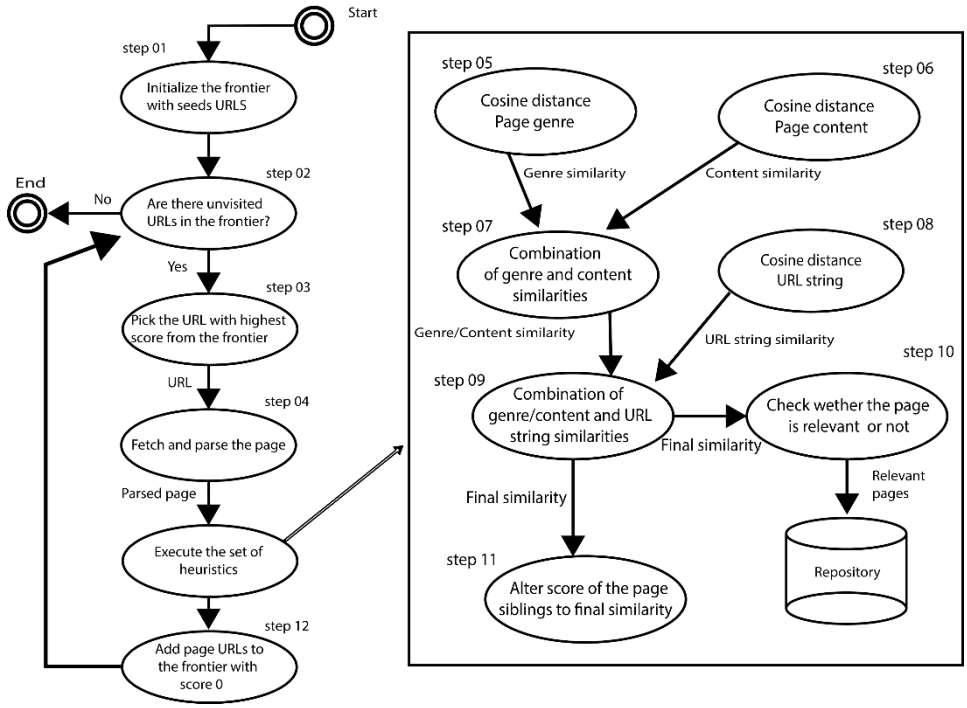

Figure 1. Functioning model of the genre-aware approach to focused crawling

Two important improvements were proposed for the genre-aware approach. The first one (Mangaravite et al., 2012) proposes the use of link-context (more precisely, anchor text, link title and URL) of the Web pages, that will be visited by the crawler, in order to define their priorities (initial scores) in the Frontier, allowing the crawler to visit the relevant pages as soon as possible; such improvement generated 100% of efficiency increase. The second one (Mangaravite et al. 2014) proposes several heuristics to semi-automatically generate seed-pages for a specific topic based on its genre and content terms: the proposed strategy (see Figure 2) generates, semi-automatically, queries to be submitted to a search engine, based on heuristics for combination of genre and content terms regarding the user interest topic. The specified heuristics for terms combination were:

- *unionOR* and *union*: heuristics that gather the genre and content terms in one single query, adding or not the logical connective OR, respectively;
- *unionFirstOR* and *unionFirst*: heuristics that only gather the first genre term and the first content term in one single query, adding or not the logical connective OR, respectively;
- *intersection* and *intersectionFirst*: heuristics that perform an intersection between either all or only the first terms of genre and content in one single query, adding the logical connective AND;
- *justContent* and *justContentOR*: heuristics that only gather the content terms in one single query, adding or not the logical connective OR, respectively;
- *justGenre* and *justGenreOR*: heuristics that only gather the genre terms in one single query, adding or not the logical connective OR, respectively.

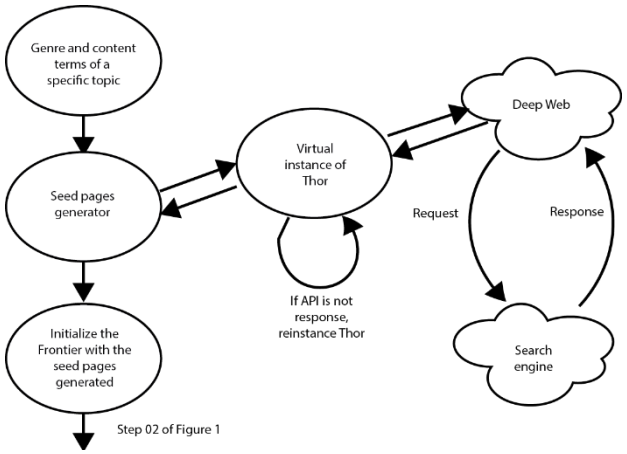

Figure 2. Functioning model of the strategy for generating seed-pages

According to the experiments, the best heuristic to semi-automatically generate seed-pages was the unionFirst that improved the efficiency of the genre-aware approach by 53%.

2.2 Automatic Threshold Estimation for Data Matching Applications

According to Santos et al. (2008), many data integration processes, such data deduplication and similarity querying, rely on the application of similarity functions. Similarity functions usually require setting a threshold value to evaluate whether a particular object is similar or not.

In order to eliminate human intervention and being capable of estimating precision and recall values for a data set, it is possible to determine similarity threshold values that tend to get the maximum F1 value. Therefore, the approach proposed in in Santos et al. (2008) performs a data clustering process aiming to, ideally, each cluster contains only instances from the same class. The main idea is trying to maximize a metric named silhouette coefficient. According to (Kaufman & Rousseeuw, 2009), the silhouette coefficient is a dimensionless quantitative metric, which can be used to determine the quality obtained by a clustering process, i.e., the similarity between elements from the same cluster and the dissimilarity between elements from different clusters.

The experiments performed in Santos et al. (2008) showed that, in 9 of 16 cases covered by the work, the curve generated by the F1 values followed the same behavior as the curved generated by the silhouette coefficient values. That is, the best silhouette coefficient value, related to a determined similarity threshold value, tends to indicate the best F1 for a given data set, as it can be seen on Figure 3. Thus, we propose a strategy based on this approach (see Subsection 3.3).

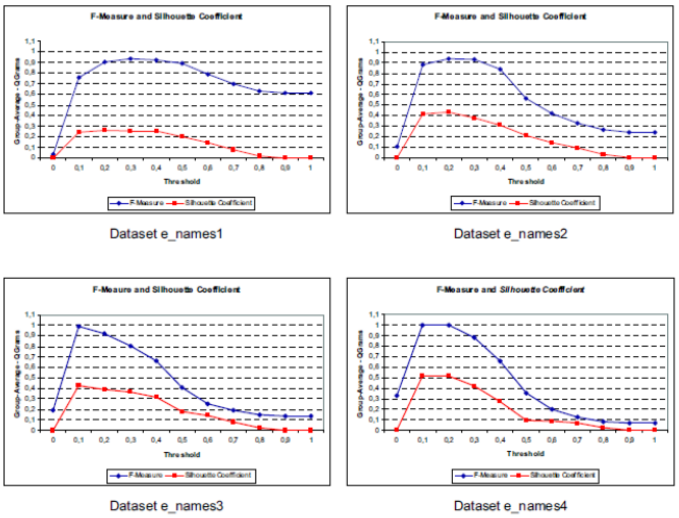

Figure 3. Results obtained by Santos et al. (2008)

3. STRATEGIES FOR DETERMINING AUTOMATICALLY SIMILARITY THRESHOLD VALUES

As already mentioned, this work has, as main goal, proposes strategies to automatically determine similarity threshold values to be used on distinct focused crawling processes based on genre-aware approach (Assis et at., 2009). The following subsections present three proposed strategies to achieve this end.

3.1 Strategy based on Seed-Pages Summarization

This strategy (see Algorithm 1) aims to determine the similarity threshold value, for a specific topic of interest, by the arithmetic or weighted mean of the seed-pages similarities according to genre and content terms from the specified topic. These pages are automatically generated by the strategy illustrated by Figure 2 also using genre and content terms from the specified topic. Figure 4 presents, diagrammatically, the architecture of the proposed strategy where genre and content terms as well as the seed-pages set directly influence on the obtained value for the similarity threshold to be used in the focused crawling process to which it is associated.

As shown in Figure 4, the proposed strategy receives, as input, the genre and content terms sets for the interest topic and the seed-pages set generated previously (see Figure 2); using these sets, it is computed the arithmetic or the weighted mean between the similarity score from the seed-pages according to the interest topic. Algorithm 1 presents formally the mentioned strategy.

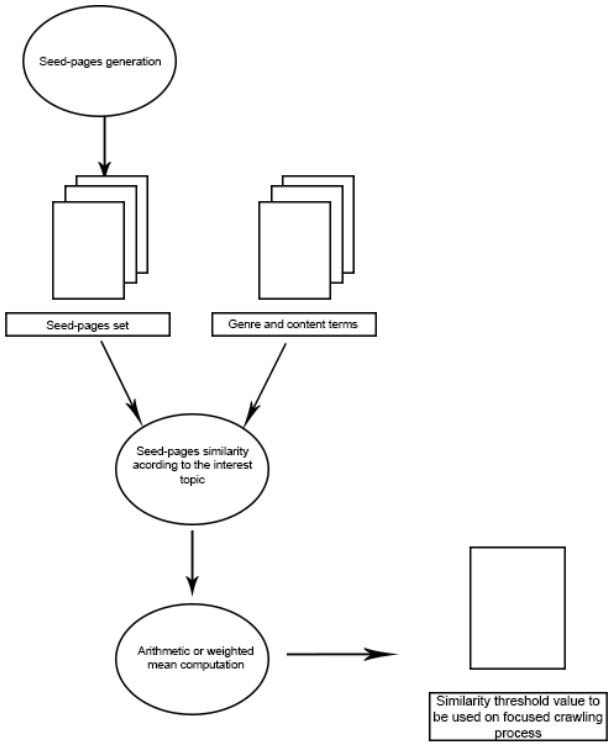

Figure 4. Strategy based on seed-pages summarization functional architecture

According to Algorithm 1, initially (steps 04, 05 and 06), for each seed-page s from set S, a similarity value is calculated between such seed-page and the genre and content terms specified, being stored in the similarities set. Following (step 07), the outliers similarity values are removed of the similarities set and, finally (step 08), the arithmetic or weighted mean between the remaining similarities is calculated. Then, the function returns (step 09) the calculated mean (a value in the range [0; 1.0]), representing the similarity threshold value to be used in a focused crawling process of the topic of interest.

Algorithm 1: Strategy based on seed-pages summarization pseudo-code
1: function GETTHRESHOLD(*S, genre, content*)
2: *threshold* ← 0
3: *similarities* ← Ø
4: **for** each seed *s* ∈ *S* **do**
5: *similarities* ← *similarities* ∪ {*similarity* (*s, genre, content*)}
6: **end for**
7: *similarities* ← *similarities – outliers* (*similarities)*
8: *threshold* ← *mean (similarities)*
9: **return** *threshold*
10: end function

Algorithm 1. Strategy based on seed-pages summarization pseudo-code

3.2 Strategy based on Clustering Methods

This strategy (see Algorithm 2) aims to determine the similarity threshold value, for a specific topic of interest, using a clustering method on the seed-pages similarities. For such strategy, we defined two distinct and classical clustering methods : K-Means (partitioning category) and BIRCH (hierarchical category).

K-Means (Hartigan & Wong, 1979) is a clustering method that requires an input parameter; in our strategy, we defined the *K* value equals to 2: one of the clusters represents the relevant pages' similarity values according to a desired topic, and the other one represents non-relevant pages' similarity values, which are disregarded because they are considered as outliers on the seed-pages' similarity values set.

BIRCH (Zhang et al., 1996) is a clustering method that requires a data instance considered in the clustering process; this data instance corresponds to the similarity value obtained between a seed-page terms and genre and content terms of the desired topic.

Figure 5 presents, diagrammatically, the architecture of the proposed strategy where genre and content terms as well as the seed-pages set directly influence on the obtained value for the similarity threshold to be used in the focused crawling process to which it is associated.

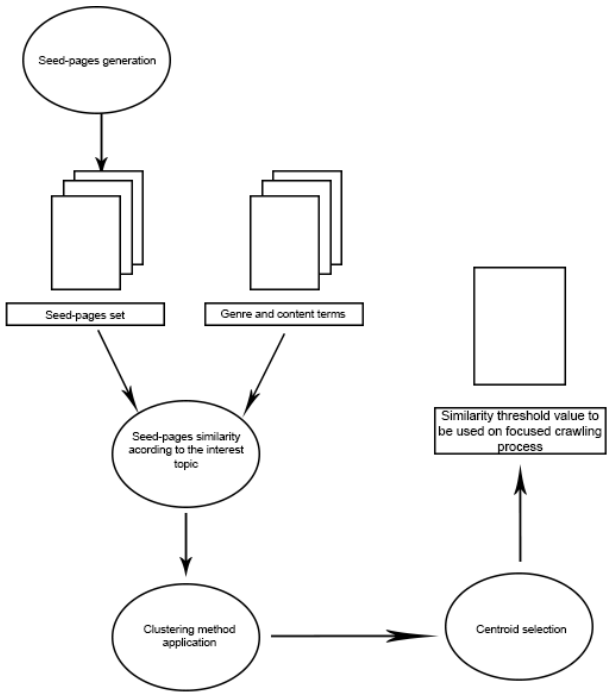

Figure 5. Strategy based on clustering methods functional architecture

As shown in Figure 5, the proposed strategy receives, as input, the genre and content terms sets for the interest topic and the seed-pages set generated previously (see Figure 2); using these sets, it is applied a clustering method on the similarities obtained between the seed-pages and the interest topic in order to obtain its centroid. Algorithm 2 presents formally the mentioned strategy.

```
Algorithm 2: Strategy based on clustering methods pseudo-code
1: function GETTHRESHOLD(S, genre, content)
2:     threshold ← 0
3:     similarities ← Ø
4:     for each seed s ∈ S do
5:         similarities ← similarities ∪ {similarity (s, genre, content)}
6:     end for
7:     clusters ← clustering (similarities)
8:     threshold ← centroid (clusters)
9:     return threshold
10: end function
```

Algorithm 2. Strategy based on clustering methods pseudo-code

According to Algorithm 2, initially (steps 04, 05 and 06), for each seed-page *s* from set *S*, a similarity value is calculated between such seed-page and the genre and content terms specified, being stored in the *similarities* set. Following (step 07), the clusters are generated according to the used clustering method and, finally (step 08), the similarity threshold value receives the centroid value of the relevant cluster generated. Then, the function returns (step 09) such centroid value (a value in the range [0; 1.0]), representing the similarity threshold value to be used in a focused crawling process of the topic of interest.

3.3 Strategy based on Silhouette Coefficient

This strategy (see Algorithm 3) aims to determine the similarity threshold value, for a specific topic of interest, by the maximization of the silhouette coefficient metric on clusters, dynamically generated, which contain relevant and non-relevant pages according to the seed-pages similarities of the desired topic. This is due to the fact that silhouette coefficient values follows the same structural behavior as the metric F1, as presented on Figure 3. Figure 6 presents, diagrammatically, the architecture of the proposed strategy where genre and content terms as well as the seed-pages directly influence on the obtained value for the similarity threshold to be used in the focused crawling process to which it is associated.

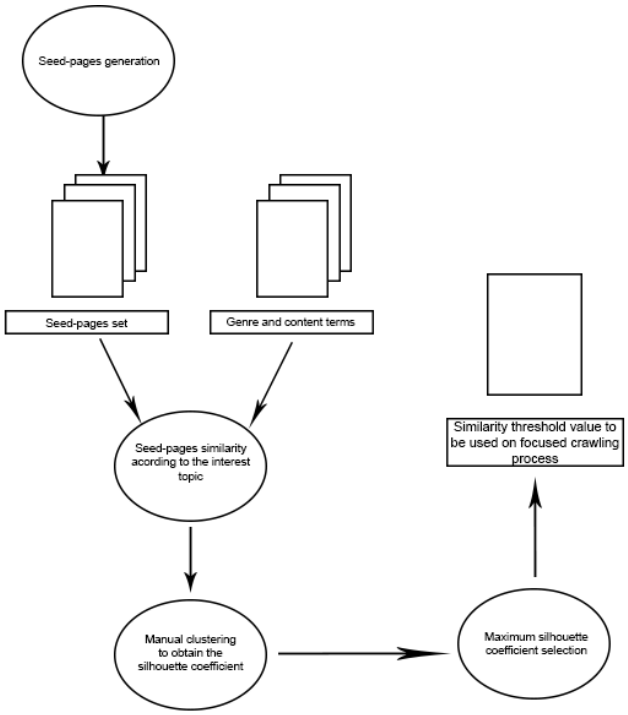

Figure 6. Strategy based on silhouette coefficient functional architecture

As shown in Figure 6, the proposed strategy receives, as input, the genre and content terms sets for the interest topic and the seed-pages set generated previously (see Figure 2); using these sets, it is performed multiple clustering processes to compute distinct silhouette coefficients in order to select the threshold associated to the maximal silhouette coefficient obtained. Algorithm 3 presents formally the mentioned strategy.

Algorithm 3: Strategy based on silhouette coefficient pseudo-code
1: function GETTHRESHOLD(*S, genre, content*)
2: *threshold* ← 0
3: *similarities* ← Ø
4: **for** each seed *s* ∈ *S* **do**
5: *similarities* ← *similarities* ∪ {*similarity* (*s, genre, content*)}
6: **end for**
7: *threshold* ← *ASTS (similarities, 0, 1.0, 0.01)*
8: **return** *threshold*
9: end function

Algorithm 3. Strategy based on silhouette coefficient pseudo-code

According to Algorithm 3, initially (steps 04, 05 and 06), for each seed-page *s* from set *S*, a similarity value is calculated between such seed-page and the genre and content terms specified, being stored in the *similarities* set. Following (step 07), the threshold receives the

return value from the function *ASTS*[1]. Then, the function returns (step 08) such calculated value by function *ASTS* (a value in the range [0; 1.0]), representing the similarity threshold value to be used in a focused crawling process of the topic of interest.

4. EXPERIMENTAL EVALUATION

In order to validate the strategies proposed on subsections 3.1, 3.2 and 3.3, we performed an experimental evaluation described on Subsection 4.1; the obtained results are described and analyzed on Subsection 4.2. In addition, Subsection 4.3 describes and analyzes an additional experiment in order to evaluate the influence of the seed pages set on the results of the strategies proposed on subsections 3.1, 3.2 and 3.3.

4.1 Experimental Setup

The experiments were conducted on the Brazilian Web, in order to crawl pages related to three topics: (1) syllabi of database courses, (2) job offers on the computing area and (3) trip to Paris. The first and the second topics were used on the practical validation experiments on the genre-aware approach to focused crawling (Mangaravite et al. 2014, Mangaravite et al. 2012, Assis et al. 2009; Assis et al. 2007), also used as basis for this paper's experimental evaluation. The third topic was defined in order to evaluate the proposed strategies on execution time.

For each topic, an expert manually specifies the set of terms required to represent the genre and the content of the desired pages. The expert also specifies the answer set containing the relevant pages among those visited by the crawler in order to allow us to assess the efficiency of the crawling processes. Table 1 shows the genre and content terms specified for each topic. We used 5 content terms and 7 genre terms for database syllabi, 6 content terms and 13 genre terms for job offers, and 10 content terms and 12 genre terms for trip to Paris.

Table 1. Genre and content terms specified for each topic

Topic	Genre terms	Content terms
Database syllabi	syllabus, course, tentative schedule, class, works, required text, textbooks	data model, relational model, entity-relationship, relational algebra, relational calculus
Job offers	jobs, job offer, vacancies, company, requirements, contact, responsibilities, location, city, description, salary, benefits, career	programmer, developer, system analyst, software, database, internet
Trip to Paris	trip, travel, accommodation, host, stay, passage, package, tour, tourist, air, land, guide	Paris, Arc de Triomphe, Eiffel Tower, Louvre, Notre-Dame Cathedral, Seine River, Musée d'Orsay, Disneyland Resort Paris, Basilica of Sacré Cœur, Panthéon

[1] An adaptation from the function proposed by Bessas et al. (2016), which computes a set of similarity threshold values and select the best one associated to the maximum silhouette coefficient value, also calculated by the function.

For the generation of the seed pages to be used in a crawling process related to a given topic, the best heuristic, proposed by Mangaravite et al. (2014), generated 20 seed-pages for database syllabi, 16 seed-pages for job offers and 15 seed-pages for trip to Paris.

For the topics database syllabi and job offers, we used the logs generated in Mangaravite et al. (2014), which include the pages visited by the crawler (45226 and 97636 for syllabi database and job offers, respectively) and the best empirical similarity threshold values (0.21 for both topics). For the topic trip to Paris, a new crawling process was performed, generating a log with 12743 pages (considered a satisfactory number to stop the crawling process); thus, we estimate the best empirical similarity threshold in order to achieve the highest F1 value (in this case, 0.502). The results obtained for each crawling process (see Table 2) were used as the baseline for this work.

Table 2. Baseline used for each topic

Topic	Similarity Threshold Value	Relevant pages	Precision[2]	Recall	F1
Database syllabi	0.210	1684	1.0	0.979	0.989
Job offers	0.210	60201	1.0	1.0	1.0
Trip to Paris	0.502	1970	0.452	0.715	0.554

4.2 Experimental Results

The proposed strategies, to automatically determine similarity threshold values, were applied to the focused crawling processes described on Subsection 4.1. The results obtained are described on Table 3.

As one can note from Table 3, the strategy that presented the best F1 values for the most topics is the strategy based on the clustering method K-Means. This occurs because K-Means achieved a similarity threshold value that is near to the empirical optimal value defined by the baseline (see Table 2). Other strategies presented satisfactory values, in some specific cases, as we can see for the strategy based on summarization using arithmetic mean for the topic database syllabi, which achieved the same best F1 value obtained by the strategy based on the clustering method K-Means.

Especially, for the topic trip to Paris, the strategies are better analyzed using precision value, once the recall value was estimated on the number of pages visited by the crawler when the crawling process were stopped manually. Therefore, the recall value obtained is an estimation of the possible final recall value for this specific topic. Consequently, the F1 value is also an estimation. For this topic, the strategy based on the coefficient silhouette obtained the best precision value. However, the precision value for K-Means is also a satisfactory value once it is better than the precision value of the baseline.

We can also verify that the similarity threshold value is directly proportional to the precision value and inversely proportional to the recall value. Thus it is possible to verify the importance to automatically determine similarity threshold values used by distinct topics on focused crawling processes.

[2] According to Manning and Schütze (1999), precision is a measure of the number of items selected correctly by a system, recall is the proportion of target items selected by a system and, finally, F1 is the harmonic mean between precision and recall.

Table 3. Results obtained by the proposed strategies for each topic

Topic	Strategy applied	Similarity threshold Value	Precision	Recall	F1
Database syllabi	Arithmetic mean	0.214	1.0	0.879	0.935
	Weighted mean	0.195	0.427	1.0	0.598
	K-means	0.214	1.0	0.879	0.935
	BIRCH	0.175	0.377	1.0	0.547
	Silhouette coefficient	0.270	1.0	0.038	0.073
Job offers	Arithmetic mean	0.277	1.0	0.055	0.010
	Weighted mean	0.278	1.0	0.054	0.102
	K-means	0.229	1.0	0.965	0.982
	BIRCH	0.310	1.0	0.008	0.102
	Silhouette coefficient	0.390	1.0	0.166×10^{-3}	0.332×10^{-3}
Trip to Paris	Arithmetic mean	0.491	0.329	0.849	0.475
	Weighted mean	0.499	0.248	0.943	0.393
	K-means	0.579	0.690	0.130	0.219
	BIRCH	1.0	0.0	0.0	0.0
	Silhouette coefficient	0.590	0.698	0.078	0.141

4.3 Seed-pages Influence

Once the strategies proposed to automatically determine similarity threshold values for focused crawling processes using the genre-aware approach (Assis et al., 2009) directly depend on the seed-pages set given as input, this Subsection has, as its main goal, evaluate the seed-pages influence on the results obtained as similarity threshold values by the strategies.

Considering the seed-pages set used on the crawling processes realized by Mangaravite et al. (2014), which were ordered by their respective similarity values according to the desired topic, we applied the following heuristics to generate subsets:

- Inter Quartile Range (IQR): selects all the seed-pages whose similarity values are between the first and the third quartile generated by considering the similarity values of the original seed-pages set;
- Random Choice: selects random seed-pages representing 25%, 50% and 75% of the original seed-pages set, considering a confidence interval of 95%;
- N-First: selects the *n* first seed-pages that represents 25%, 50% and 75% of the original seed-pages set;
- N-Last: selects the *n* last seed-pages that represents 25%, 50% and 75% of the original seed-pages set;
- Full-set: selects the original seed-pages set completely.

The proposed heuristics were evaluated using the first proposed strategy to automatically determine similarity threshold values (see Subsection 3.1) due to the previous knowledge of the sensibility in computing the mean when the original seed-pages set suffers modifications. The results obtained, considering the topics database syllabi and job offers, are shown in Table 4.

Table 4. Seed-pages influence on similarity threshold values generated

	Database syllabi		**Job offers**	
	Arithmetic mean	**Weighted mean**	**Arithmetic mean**	**Weighted mean**
IQR	0,214	0,195	0,276	0,251
Random Choice (25%)	(0,142; 0,233)	(0,197; 0,239)	(0,251; 0,299)*	(0,204; 0,321)
Random Choice (50%)	(0,177; 0,223)	(0,164; 0,221)	(0,258; 0,283)*	(0,256; 0,324)*
Random Choice (75%)	(0,195; 0,214)	(0,169; 0,210)	(0,262; 0,298)*	(0,239; 0,288)*
N-First (25%)	0,202	0,197	0,329	0,333
N-First (50%)	0,201	0,197	0,277	0,299
N-First (75%)	0,194	0,193	0,260	0,283
N-Last (25%)	0,189	0,189	0,264	0,250
N-Last (50%)	0,190	0,192	0,276	0,260
N-Last (75%)	0,196	0,218	0,288	0,259
Full-set	0,192	0,195	0,277	0,278

The results on Table 4 showed that the original seed-pages set and the form the data is organized, according to the desired topic, influence on the final similarity threshold value obtained by the analyzed strategy. It is also possible to verify that in the cases where the seed-pages set is extremely similar to the desired topic, the result obtained for the similarity threshold value was higher than the ideal value. Moreover and particularly, using the Random Choice heuristic, some executed instances presented, as similarity threshold values interval, a range that did not contain the optimal similarity threshold; these instances are marked by the (*).

In general, in order to obtain results for similarity threshold values that are closer to the optimal defined previously, the best heuristic was the IQR. This heuristic presented as result, using Arithmetic Mean, the same similarity threshold value presented by the strategy based on clustering method K-Means (see Table 3).

5. CONCLUSION

As presented in Section 4, evaluation experiments for the proposed strategies for automatic determination of similarity threshold values to be used in focused crawling processes, based on genre-aware approach, were realized obtaining, as the best result, 100% of precision and 98% of F1 considering the topic job offers (see Subsection 4.1). This strategy also achieved the best results for the other topics in analysis, which make it the best strategy proposed in this work.

As future work, we intent to: (1) propose a new architecture for the crawler where it is possible to perform a distributed focused crawling process; (2) use natural language processing to improve the genre and content terms defined for the desired topic of interest and, thus, to improve the vector model evaluation of the crawling approach; (3) evaluate proposed strategies on different focused crawling approaches.

ACKNOWLEDGEMENT

This research was partially funded by research grants from PIP/UFOP and by project MASWeb (grant FAPEMIG/PRONEX APQ-01400-14). Furthermore, it was carried out on the GAID/UFOP Laboratory.

REFERENCES

Almpanidis, George, and Constantine Kotropoulos, 2005, Combining text and link analysis for focused crawling. *Pattern Recognition and Data Mining. Springer.* Berlin Heidelberg, pp. 278-287.

Assis, G.T. de et al., 2009, A Genre-Aware Approach to Focused Crawling. *World Wide Web*, 12(3), pp. 285–319.

Assis, G.T. de et al., 2007. Exploiting genre in focused crawling. *In Proceedings of the 14th International String Processing and Information Retrieval*. Santiago, Chile, pp. 62–73.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto, 1999. *Modern information retrieval*. Vol. 463. New York: ACM press.

Bessas, Izaquiel L., et al., 2016, Automatic and online setting of similarity thresholds in content-based visual information retrieval problems. *EURASIP Journal on Advances in Signal Processing 2016*.1, pp. 1-16.

Chakrabarti, Soumen, Martin Van den Berg, and Byron Dom., 1999, Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks 31.11* pp. 1623-1640.

Hartigan, John A., and Manchek A. Wong. 1979, Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1, pp.* 100-108.

Santos, Juliana Bonato, et al., 2008, Automatic threshold estimation for data matching applications.*Proceedings of the 23rd Brazilian symposium on Databases. Sociedade Brasileira de Computação.*

Johnson, Judy, Kostas Tsioutsiouliklis, and C. Lee Giles, 2003, Evolving strategies for focused web crawling. *ICML*.

Kaufman, L., & Rousseeuw, P. J., 2009. *Finding groups in data: an introduction to cluster analysis (Vol. 344)*. John Wiley & Sons.

Mangaravite, V., Assis, G.T. De & Ferreira, A.A., 2012. Improving the Efficiency of a Genre-aware Approach to Focused Crawling Based on Link Context. *In 8th Latin American Web Congress*. pp. 17–23.

Mangaravite, V., Assis, G.T. De & Ferreira, A.A., 2014. Semi-automatic generation of seed pages in genre-aware focused crawling. *In Proceedings of the 13th International Conference WWW/Internet (ICWI)*. pp. 51-58.

Manning, C. D., & Schütze, H., 1999. *Foundations of statistical natural language processing (Vol. 999)*. Cambridge: MIT press.

Menczer, F., Pant, G. & Srinivasan, P., 2004, Topical Web Crawlers: Evaluating Adaptive Algorithms. *ACM Transactions on Internet Technology, 4(4)*, pp. 378–419.

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

Pant, G., & Srinivasan, P. (2006). Link contexts in classifier-guided topical crawlers. *Knowledge and Data Engineering, IEEE Transactions on, 18(1)*, pp. 107-122.

Pant, Gautam, et al., 2004, Panorama: extending digital libraries with topical crawlers. *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries. ACM*.

Zhang, Tian, Raghu Ramakrishnan, and Miron Livny., 1996, BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record*. Vol. 25. No. 2. ACM.