A General Evaluation Framework for Adaptive Focused Crawlers

Fabio Gasparetti, Alessandro Micarelli and Giuseppe Sansonetti
Department of Engineering, Roma Tre University, Via della Vasca Navale 79, Rome, Italy

Keywords: Adaptive Focused Crawling, Evaluation Framework.

Abstract: Focused crawling is increasingly seen as a solution to increase the freshness and coverage of local repository of documents related to specific topics by selectively traversing paths on the web. The adaptation is a peculiar feature that makes it possible to modify the search strategies according to the particular environment, its alterations and its relationships with the given input parameters during the search. This paper introduces a general evaluation framework for adaptive focused crawlers.

1 INTRODUCTION

Due to the limited bandwidth, storage and resources of traditional computational systems and the rapid growth of the web, focused crawlers aim at building small high-quality and up-to-date repositories of topic-specific pages. Deep analyses of the retrieved pages have also the chance to better address growing dynamic contents, such as news or financial data and promptly alerting about relevant alterations of the retrieved pages.

Pages related to the same topics tend to be neighbours of each other is the fundamental assumption that is often named *topic locality* (Davison, 2000). Thus, the objective of the crawlers is to stay focused, that is, remaining within the neighbourhood in which topic-specific pages have been identified.

In this context, an evaluation methodology is a logical description of the processes and connected elements to be followed to help one better understand a quality evaluation. By following this process, a computer scientist or practitioner can learn what he or she needs to know to determine the level of a performance of a search strategy in a specific context. This paper is geared toward a definition of a evaluation methodology for the adaptive focused crawlers.

At present, focused crawling evaluations that also include adaptivity analysis are not available. One of the reasons could be the difficulty to measure the reaction of crawlers to user needs refinements or alterations of the environment. How long does it take to adapt the crawl to a user relevance feedback and provide new interesting documents? How many environment alterations are tolerable before the crawling performance falls below a given threshold? Standard methodologies to assess those characteristics, thus allowing comparing different search strategies are yet to be developed.

The aim of this paper is twofold. First, we summarise the different evaluation approaches that have been proposed in the literature, critically discussing the testbed settings and the evaluation metrics. After having identified the most relevant factors to be included in a general framework, we give an account of the fundamental elements for the definition of an evaluation methodology regarding *adaptive* focused crawling systems.

The paper is organized as follows. We first present the most relevant input data that distinguish a specific evaluation in Sections 3.1 and 3.2. Section 3.3 deals with the approaches proposed for the definition of relevance measures of the retrieved pages. In Section 3.4, we consider the resource constraints, while Sect. 3.5 is focused on the measures that better characterise the effectiveness of the search strategies during the crawl. Section 3.6 introduces the evaluation approaches based on comparative analysis, while 3.7 specifically dwells on the assessment of the adaptive behaviour of the crawlers. The following section discusses the related work in the literature. The last section is a conclusion.

2 RELATED WORK

The foremost exploratory research activity on the evaluation of adaptive focused crawlers has been proposed by Menczer *et al.* in (Menczer et al., 2004).

Gasparetti F., Micarelli A. and Sansonetti G..
A General Evaluation Framework for Adaptive Focused Crawlers.
DOI: 10.5220/0004965903500358
In *Proceedings of the 10th International Conference on Web Information Systems and Technologies* (WEBIST-2014), pages 350-358
ISBN: 978-989-758-024-6

In particular, they compare several crawlers based on machine learning techniques in order to assess both the obtained general performance and some characteristics of adaptivity. The authors' principal goal was to evaluate the benefits of the machine learning versus other approaches. While machine learning has the chance to play a key role in the development of focused crawlers able to automatically adapt their search strategies to the peculiar characteristics of the topics and environment, the proposed framework misses to cover scenarios when the approaches are subjected to continuous updates in the input data (i.e., topics and environment alterations). In this case, adaptivity can be performed either incrementally by continuous update or by retraining using recent batches of data, either new or already visited pages subjected to updates. In this scenario, the relation between the input data and the target variable changes over time.

Several other frameworks have been proposed (Menczer et al., 2001; Chau and Chen, 2003; Srinivasan et al., 2005; Pant and Srinivasan, 2005), but none of them explicitly include adaptive behaviour analysis.

3 AN EVALUATION FRAMEWORK FOR ADAPTIVE FOCUSED CRAWLERS

Defining an evaluation methodology for a *standard* crawler does not require a great effort. Once a subset of the web is available, it is possible to run an instance of the crawl on a workstation and monitor the most important parameters to measure its effectiveness (Cho et al., 1998). The proactivity and autonomy characteristics of the search strategies of focused crawlers, which potentially allow them to explore regions of the web far from the starting points, call for different evaluation approaches.

In addition to that, if the focused crawlers have some sort of adaptivity behaviour w.r.t changes in the environment or the current topics, the evaluation framework should keep track of changes of the performances and behaviour when one of both of these aspects are being altered. Good adaptivity is characterised by changes of unconstructive or disruptive behaviour, often caused by external stimuli, to something more constructive, which is able to fulfil the goal of the search activity.

In the following sections we define the parameters and the most relevant elements that form the evaluation methodology, to be assessed and reported during the experiments with adaptive focused crawlers.

3.1 Corpus

There are two broad classes of evaluations, system evaluations and user-centred evaluations. The latter measure the user's satisfaction with the system, while the former focuses on how well the system is able to retrieve and rank documents. Several researchers accept that evaluators of adaptive systems should adopt a user-centred evaluation approach because users are both the main source of information and the main target of the application, but manually finding all the relevant documents in a large collections of billion of documents is not practical. User-based evaluation is extremely expensive and difficult to do correctly. A proper designed user-based evaluation must use a sufficiently large, representative sample of potential topics. Such considerations lead researchers to use the less expensive system evaluations.

Of course, technical issues must be addressed in order to construct a collection that is a good sample of the web (Bailey et al., 2003). Nevertheless, "bold" focused crawlers have the chance to take decisions on many different paths and visit pages far from the seed sets, with more chances to end up towards paths of pages not being included in the initial collection. For this reason, standard or predefined collections are rarely employed.

All, or almost all, of the focused crawling evaluations in the literature do not employ any corpus but allow the crawlers to access any document on the web. Web pages continue to change even after they are initially published by their authors and, consequently, it is almost impossible to make comparisons from results obtained by different search strategies, as discussed in Sect. 3.6.

The adaptivity behaviour allows crawlers to dynamically adjust the search strategies to several different and unexpected external alterations. Its evaluation is therefore a complex activity going through the identification and assessment of several variables, sometimes in mutual relationship one another. It is reasonable that a sound evaluation has to consider complex and long-lasting test evaluations to identify those relationships as a function of controlled variations in the input data. A static and large corpus of web documents is the only requirement that guarantees the valid comparison of several outcomes obtained at different times.

3.2 Seeds

A good selection of seed pages guarantees that

enough pages from different communities related to the current topic will be sampled and the crawler exploits the topical locality for finding additional pages in comparison with crawl starting from random seeds. For instance, Daneshpajouh *et al.* (Daneshpajouh et al., 2008) compared various community-based algorithms for discovering good seeds from previously crawled web graphs and discovered that HITS-based ranking is a good approach for this task. Of course, the seed page identification should not be too expensive in terms of computational time. If web corpora are not available, valid sources of seeds may be human-generated directories such as Open Directory Project (ODP)[1], where each category contains links to pages about similar topics.

Of course, seed pages related to the interesting topics make the search for related pages much easier because of the topical locality phenomenon. Srinivasan *et al.* (Srinivasan et al., 2005) provide an interesting mechanism to control the level of difficulty of the crawl tasks by means of the hypertext structure among pages. Once a subset of target pages, that is pages relevant to a topic, is identified, it is possible to collect pages linking to the specified targets by querying one of the online services such as Mozscape[2]. By iterating this *backlink* procedure, it is possible to collect several paths, a subset of them bringing to the target pages. The last pages to be collected are the ones that will be included in the seed set.

The number of iterations I match the level of difficulty of the crawling task. Particularly difficult tasks have a few relevant pages far away form the seed sets. If the crawler is able to find those targets, its edge search strategy has boldness traits favouring the exploration on various different paths. The opposite behaviour of the greedy strategies encourages the exploitation of the current good pages sticking the exploration to their vicinity. An adaptive selection of bold and greedy strategies may rely on the current acquired evidence. For example, once a number of relevant websites have been found, the exploration can be focused on the near linked pages, while bold strategies are valid when no evidence is fruitful and new paths have to be verified. At present, focused crawlers do not explicitly include this form of search strategy adaptivity.

The above-mentioned backlink procedure is the only one that allows the framework to include the recall measure of performance discussed in Sect. 3.5.1. As a matter of fact, the procedure builds up a small corpus of pages, where the good ones are clearly identified.

[1]http://www.dmoz.org

[2]http://moz.com

3.3 Topic Affinity

Ideal focused crawlers retrieve the highest number of relevant pages while simultaneously traversing the minimal number of irrelevant pages. This leads to significant savings in hardware and network resources, and helps keep the crawl more up-to-date.

One of the first evaluation parameters to take into consideration is the soundness of the retrieved documents' content. The traditional crawlers' goal is to download as many resources as possible, whereas a focused crawler should be able to filter out the documents that are not deemed related to the topics of interest. Focused crawlers respond to the particular information need expressed by topical queries or interest profiles.

Besides monitoring the exploration results, the evaluation of the relatedness of the retrieved documents is also fundamental for the selection of the best routes to follow. For this reason, focused crawlers routinely compute these measures for assigning the priorities to the queued URLs during the exploration. A formal description for the topic of interests is fundamental for effectively driving the crawling to a subset of paths and, of course, it is strictly correlated to the definition of the relatedness measure. Singular domains may also define ad-hoc measures of effectiveness, such as novelty and diversity of page contents (Barbosa and Bangalore,).

In the following sections, we give an account of the most relevant approaches for evaluating the relatedness of the retrieved documents.

3.3.1 Topic Selection

Information searching is the intent that usually motivates the queries driving the focused crawling. Users are willing to locate documents and multimedia concerning a particular topic in order to address an information need or perform a fact-finding or general exploratory activity. These topics can be along a spectrum from very precise to very vague.

A long-lasting research activity aiming at defining a comprehensive classification of user intents for web searching (e.g., (Jansen et al., 2008)) and related IR evaluations (e.g., (Sakai, 2012)) is largely available.

In contrast to search engines, topics submitted to focused crawling are defined by expert users able to accurately select a good representation of their intents. At the same time, those intents can still assume both a broad (e.g., "Find information about Windows 9") or specific scopes (e.g., "Find stocks of DNA sequencing companies").

While automatic approaches to select broad-topic queries are available (see Sect. 3.3.3), specific scope

queries are usually human-generated or extracted from real scenarios (Gasparetti et al., 2014). In spite of that, general evaluation frameworks should take into account both of the typologies in order to assess the benefits of different strategies and adaptivity techniques in the two scenarios.

3.3.2 "Plain" Matching

Several focused crawlers use text similarity measures for comparing the content extracted from the crawled pages and a representation of the topic that drives the search.

If both topics and contents are described by keywords, the relevance between them can be calculated by one of the well-known approaches proposed in the IR, such as:

VSM. Vector Space Model (e.g., (Hersovicia et al., 1998))

NB. Naive Bayes classifiers trained on a subset of documents related to the topic (e.g, (Chakrabarti et al., 1999; Chakrabarti et al., 2002))

SVM. Support Vector Machine (e.g., (Ehrig and Maedche, 2003; Choi et al., 2005; Luong et al., 2009))

NN. Neural networks (e.g., (Menczer and Monge, 1999; Chau and Chen, 2003))

LSI. Latent Semantic Indexing (e.g., (Hao et al., 2011))

The output is usually any real number between 0 and 1:

$$f_m : DxT \rightarrow [0,1]. \quad (1)$$

where D and T are the representations of the document and topic, respectively.

A comparative evaluation shows how NB classifiers are weak choices for guiding a focused crawler when compared with SVM or NN (Pant and Srinivasan, 2005).

3.3.3 Taxonomy-based Matching

In order to more accurately drive the crawl, some focused crawlers use hierarchical structures for classifying pages (Chakrabarti et al., 1999; Chen et al., 2008). There are several complex hierarchical taxonomies and ontologies available, e.g., Medical Subject Headings, U.S. Patents, ODP and CIDOC Conceptual Reference Model for cultural heritage. Instead of binary classifiers, where each category or class is treated separately, hierarchical categorisation may drop a document into none, one, or more than one category. Users instantiate a crawl by selecting one or more topics in the given taxonomy.

Imagine a hierarchy with two top-level categories, e.g., `Computers and Recreation`, and several subcategories, such as `Computers/Hacking`, `Computers/Software` and `Computers/Emulators`. In a non-hierarchical model, a word like *computer* is not very discriminating since it is associated with several categories related to computers. In a hierarchical model, more specialized words could be used as features within the top-level `Computer` category to better choose the right one for a given a document.

Chakrabarti *et al.* (Chakrabarti et al., 1999) determine the relevance of one page analysing its ancestor categories. If one of those ancestors is in the subset of topics selected by the user, the page is further analysed because it covers more detailed topics. The same approach can be employed in an evaluation framework so that relevant documents are not ignored because they do not ideally match the user topic.

Text descriptions of the descendants in the taxonomy can be used to improve the representation of the topic of interests (Chen et al., 2008). Cross-language hierarchical taxonomies can also be employed to allow focused crawlers analyse pages in different languages.

Menczer *et al.* propose to use the ODP taxonomy to automatically generate topics for the evaluations (Menczer et al., 2004). Leaves with five or more links are extracted and used as potential topics. In particular, the text corresponding to the title of the category and the anchors of the external links become a text description of each topic.

Hierarchical categorisation with SVM has been proven to be an efficient and effective technique for the classification of web content (Dumais and Chen, 2000). Other relevant approaches are based on the semantic analysis of the content, e.g., (Limongelli et al., 2011; Gentili et al., 2003; Biancalana et al., 2013).

3.3.4 Predicate-based Matching

A focused crawler estimates the likelihood that each candidate link will lead to further relevant content. Evidence such as links' anchor text, URL words and source page relevance are typically exploited in estimating link value comparing the text against the current topic of interest.

Aggarawal *et al.* (Aggarwal et al., 2001) propose the definition of arbitrary predicates in order to better perform the resource discovery. Besides simple keywords, predicates may extend to peculiar characteristics of the retrieved pages or properties of the linkage topology. By analysing the characteristics of the collected pages and the values of their predicates, it is possible to understand the statistical relationship between the predicates and the best candidate pages.

For instance, Diligenti *et al.* (Diligenti et al., 2000) use the context-graph idea to learn the characteristics of the best routes examining features collected from paths leading up to the relevant nodes.

Besides sets of keywords, predicates give users the chance to represent the features that the retrieved pages must own in order to be judged relevant. For example, opinion and discourse analysis on contents spread out on a sequence of connected pages might unveil valuable information that strict keyword-based relevance measures on single documents might miss.

While predicates are shown to be fundamental improvements in developing adaptive focused crawlers (Micarelli and Gasparetti, 2007), there are not attempts to use user-defined predicates to evaluate the performance of the crawlers. Predicates are usually very context-dependent, therefore they are strongly affected by the specific goal, situation, domain, task or problem under examination. None of the predicate-based approaches proposed in the literature propose a formal methodology for the definition of those predicates. User-defined predicates are subjective by nature, for this reason they are less suitable for being included in general evaluation frameworks.

3.3.5 Authoritativeness

The overwhelming amount of information on the web related to a potential topic of interest may hinder the ability to make important decisions. One of the advantages of focused crawlers, that is the reduction of the information overload, is only partially achieved when the topics of interest is too general or vague.

Focused crawlers use topic distillation for finding good *hubs*, i.e., pages containing large numbers of links to relevant pages for updating the current queue of URLs to visit (Kleinberg, 1998; Chakrabarti et al., 1999). The purpose of topic distillation is to increase the precision of the crawl, even if there is no trace of the topic keywords in them. Pages and links form a graph structure and connectivity analysis algorithms based on a mutual reinforcement approach is able to extract hubs and authority pages, that is relevant pages pointed by hubs. Different topics may show different topologies of interconnections between web pages. Menczer *et al.* (Menczer et al., 2004) state how iterative algorithms such as HITS and PageRank able to extract meaning from link topology permit the search to adapt to different kinds of topics.

While focused crawlers use hubs for finding new seeds during the crawl, authority measures can be used to evaluate the importance of the retrieved pages. Despite similar performances, different focused crawlers may cover subspaces on the web with low overlap. Due to dissimilar topologies, authority measures better unveil different outputs and search strategies.

A clear limitation of these measures in an evaluation framework is that they are computed on a partial web graph built by extracting the links from the collected documents obtaining a rough approximation of their values. The use of a static large corpus (e.g., CommonCrawl) can overcome this obstacle.

3.4 Resource Contraints

Focused crawling identifies relevant documents reducing the computational resources required by this task. The principal computational resources are computation time, network bandwidth and memory space. While the processing speed and memory capacity cost unit have been constantly reduced in recent years, network bandwidth poses strong limits on the number of documents that can be downloaded and evaluated.

Focused crawlers based on iterative algorithms such as HITS, e.g., (Cho et al., 1998; Chakrabarti et al., 1999; Rungsawang and Angkawattanawit, 2005) are expected to reduce the rate of page downloads when the set of hypertext documents is large. Most of the current evaluations ignore experiments that extend over 10 thousands of documents and hence they just ignore this issue. Comparative analysis of focused crawlers that include iterative algorithms should clearly state the asymptotic estimates of the complexity, therefore ignoring the efficacy alteration due to potential different implementations of the same algorithms.

In practice, a simple heuristic to determine the CPU usage is monitoring the time elapsed before reaching a given limit of retrieved documents. Results should generally be averaged over several tests and statistical significance values have to be computed in order to reduce the effects of temporary Internet slowdowns and prove the soundness of the evaluation.

3.5 Behaviour Analysis

An evaluation framework of focused crawling strategies has to provide provable guarantees about their performance assessments. However, an algorithm that works well in one practical domain might perform poorly in another. Trend analysis on each topic based on the information accumulated over a period of crawl activity would permit to understand the variations of the performances as a function of explicit or implicit variables. The complexity of the topic, the amount of links in the visited web graph or the unreachable pages are only some of the variables that may strongly alter the behaviour of the focused crawlers. While

an average on several tests may reduce the influence of these variables on the final results, analysing some measures during the crawl gives the chance to get different views on the performance and better characterise the benefits and drawbacks of various strategies in various contexts.

3.5.1 Precision, Recall and Harvest Rate

Precision and recall are two popular performance measures defined in automated IR, well defined for sets. The former P_r corresponds to the fraction of top r ranked documents that are relevant to the query over the total number of retrieved documents, interesting and not.

$$P_r = \frac{found}{found + false\ alarm} \quad (2)$$

while recall R_r is the proportion of the total number of relevant documents retrieved in the top r (cutoff) over the total number of relevant documents available in the environment:

$$R_r = \frac{found}{found + miss} \quad (3)$$

Precision and recall are virtually independent by the relatedness measure definition, therefore it is possible to employ one of the above-mentioned measures in order to identify relevant and irrelevant documents.

As pointed out in (Chakrabarti et al., 1999), the recall indicator is hard to measure because it is impossible to clearly derive the total number of documents relevant to a topic due to the vastness of the web, unless the backlink procedure for the seed selection discussed in Sect. 3.2 is chosen.

If the precision of the fetched pages is computed during the crawl, the curve of *harvest rate* $h_r(n)$ for different time slices of the crawl is obtained, where n is the current number of fetched pages (Chakrabarti et al., 1999). This measure indicates if the crawler gets lost during the search or if it is able to constantly keep the search over the relevant documents. The harvest rate becomes a critical measures for analysing the behaviour of the crawlers after alterations of the environment or topics of interests (see Sect.3.7).

3.5.2 Deep Web Strategies

On a different note, most search engines cover what is referred to as the publicly indexable Web but a large portion of the Internet is dynamically generated and such content typically requires users to have prior authorisation, fill out forms, or register (Raghavan and Garcia-Molina, 2001). Other information refers to Twitter or Facebook posts, links buried many layers down in a structured website, or results that sit so far down the standard search results that typical users will never find them. This covert side of the Internet is commonly referred to as the hidden/deep/invisible web. Hidden web content often corresponds to precious information stored in specialised databases. Focused crawlers have the chance to include novel deep web crawling strategies in order to find out additional relevant documents (Zheng et al., 2013). A feasible measure to assess the effectiveness of these strategies is based on the comparison of the retrieved pages with the collection of pages retrieved by popular search engines. A large subset of relevant documents that does not overlap with the search engines' collections is expected for good deep web strategies.

While these strategies are golden features useful in several contexts and, therefore, required to be evaluated during the crawl, a very few attempts have been proposed, and all of them limit the scope of their techniques to strategies for specific portions of websites (Bergholz and Chidlovskii, 2003; Liakos and Ntoulas, 2012).

3.6 Comparative Analyses

Section 3.1 discussed how focused crawling can be seen as a particular instance of the IR task, which goal is selecting a subset of documents from a large collection relevant to a given topic. For this reason, the long-lasting research activity in the IR evaluation has the chance to support new frameworks for focused crawlers.

Several IR experiments are designed following the Cranfield paradigm, where same sets of documents, topics and measures are used for various approaches that are considered in isolation, freed as far as possible from the contamination of operational variables. The experimental design calls for same corpus of hypertext documents and same topics, with the computation of the same effectiveness measures in order to directly compare different approaches' outcomes. A performance comparison between adaptive, non-adaptive and unfocused crawlers (e.g., breadth-first, random strategy) can be easily obtained.

While hypertext test collections have been often used in the IR domain (e.g., the ones provided by the Text Retrieval Conference TREC), they show several drawbacks in the focused crawling as discussed in Par.3.1.

Current focused crawling evaluations (e.g., (Srinivasan et al., 2005; Menczer et al., 2004)) follow a hybrid approach, where each round of tests are based on the same topic and measures but each single strategy is evaluated allowing the crawler to directly access the

web. The authors make the assumption that the web is not being altered between two evaluations. Except if the evaluations take place very quickly one after another, this assumption is clearly wrong.

A partial workaround consists in caching the accessed pages so that future requests and evaluations for that data can be served ignoring potential occurred alterations. Due to the locality reference of crawlers, it is also possible to cache pages that are connected by the ones that have been retrieved during the crawl (i.e., prefetching). While caching can simulate similar testbeds between evaluations of different search strategies, it fails to maintain consistency between the cache's intermediate storage and the location where the data resides. For example, home pages of news websites such as CNN.com are usually altered several times a day while other sections are not. Caching techniques that store only part of these websites cannot reproduce a valid image of their hypertext structure and reachable content. Once again, a large snapshot of the web is the only feasible way to guarantee the same platform for experimentation for various search strategies.

3.7 Adaptivity

Menczer *et al.* associate *adaptivity* to the approaches that include any sort of machine learning techniques for guiding search (Menczer et al., 2004). Adaptive techniques are basically seen as means to better understand the environment and its peculiar relationships with the topic. The environment and the topics are perceived as static features.

On a different note, Micarelli and Gasparetti (Micarelli and Gasparetti, 2007) extend the definition of adaptive focused crawlers to the ones able to address potential variations in the environment or in the topic definition. As a matter of fact, two relevant adaptive crawlers (Menczer and Monge, 1999; Gasparetti and Micarelli, 2003) show both adaptive behaviour implementing multi-agent evolutionary or optimisation algorithms. For this reason, we should like to propose a methodology and measures to effectively assess this form of adaptation, whatever technology is chosen for the implementation of the focused crawlers.

3.7.1 Domain Adaptivity

Empirical analysis of web page changes combined with estimates of the size of the web states how an mount close to 5% of the indexable web must be subjected to re-index daily by search engines to keep the collection up-to-date (Brewington and Cybenko, 2000). Several statistical approaches aim at predicting whether a page will change based on the change frequency for the same page observed over some past historical window (Radinsky and Bennett, 2013). Nevertheless to our knowledge, there is not any focused crawler that implements a scheduling policy for revisiting web pages and adaptively alters its search strategy accordingly. Tight restrictions on the network bandwidth do not allow to allocate enough resources for collecting evidence about change rates of pages. Without these data, robust computation of temporal change patterns and prediction activity are not possible.

Singular exceptions are focused crawlers based on genetic or ant paradigm approaches (Gasparetti and Micarelli, 2003; Menczer et al., 2004). In both the approaches, a population of autonomous agents are able to visit the environment collecting evidence about potential alterations of content and hypertext structure. In spite of that, the authors have not included the domain adaptivity characteristic in the evaluation of their approaches, nor have not future comparative studies done.

At the same time, estimating the importance of each page during the discovery of the web graph is one of the goals of the focused crawlers (Abiteboul et al., 2003). In some circumstances, such as stock market news, the importance is affected by the freshness of the published content. Efficient focused crawling strategies call for a better understanding of this relationship and adaptively change the behaviour of search for uncover the largest number of important resources.

A feasible approach for evaluating these aspects is by empirically measuring the time requested to revisit a carefully defined set of pages that have been subjected of alterations in their content. In particular, once a large set of cached pages have been collected by previous evaluations, it is possible to identify the subset W_a of these pages that have been altered since the beginning of the tests. By randomly choosing a subset $W'_a \subseteq W_a$ that satisfies a given percentage of unique changes (e.g., 5%) and, at the same time, is related to the current topics, new evaluations are performed. Good adaptive strategies will access to these pages W'_a sooner keeping the available computational resources the same (see Par. 3.4).

3.7.2 Topic Adaptivity

A more subtle form of adaptivity regards the topic that guides the crawl. Queries or interest profiles might be altered during the crawl in various ways:

- Generalisation: A similar or new topic seeking more general information than the previous one;
- Specialisation: A similar or new topic seeking

more specific information than the previous one;

- Reformulation: A new topic that can be viewed as neither a generalisation nor a specialisation, but a reformulation of the topic.

While traditional IR approaches consider each query independently one another, focused crawlers have the chance to exploit collected evidence during previous crawls to drive future exploratory activities on similar topics saving computational resources.

Generalisation and specialisation are two forms of topic alterations that can be easily automated by employing a taxonomy-based representation of topics, as discussed in Sect. 3.3.3. Lower levels of these forms of topic organisations correspond to specialisation while upper levels to generalisation. During comparative analysis, several search strategies may be affected by the same topic alteration. By monitoring the impact of this alteration on the performance measures (e.g., average topic affinity of pages) it is possible to identify the approaches that better exploit the collected evidence being able to promptly adapt the exploration.

4 CONCLUSIONS

The major contribution of the present position paper is to propose an extended evaluation framework for focused crawlers able to take into consideration adaptivity behaviours. A developed discussion on the limitations of the current approaches allowed us to identify relevant features that have currently been ignored in the literature. Moreover, using the lessons learned from the previous crawler evaluation studies, the proposed framework makes explicit reference to the measures proven to be fundamental so far.

We are currently planning to apply the described methodology in a real scenario, where a comparative analysis will analyse the performance of the most popular adaptive focused crawlers.

REFERENCES

Abiteboul, S., Preda, M., and Cobena, G. (2003). Adaptive on-line page importance computation. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 280–290, New York, NY, USA. ACM.

Aggarwal, C. C., Al-Garawi, F., and Yu, P. S. (2001). Intelligent crawling on the world wide web with arbitrary predicates. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 96–105, New York, NY, USA. ACM.

Bailey, P., Craswell, N., and Hawking, D. (2003). Engineering a multi-purpose test collection for web retrieval experiments. *Inf. Process. Manage.*, 39(6):853–871.

Barbosa, L. and Bangalore, S. In Macdonald, C., Ounis, I., and Ruthven, I., editors, *CIKM*, pages 755–764. ACM.

Bergholz, A. and Chidlovskii, B. (2003). Crawling for domain-specific hidden web resources. In *Proceedings of the Fourth International Conference on Web Information Systems Engineering*, WISE '03, pages 125–, Washington, DC, USA. IEEE Computer Society.

Biancalana, C., Gasparetti, F., Micarelli, A., and Sansonetti, G. (2013). Social semantic query expansion. *ACM Trans. Intell. Syst. Technol.*, 4(4):60:1–60:43.

Brewington, B. E. and Cybenko, G. (2000). How dynamic is the web? In *Proceedings of the 9th International World Wide Web Conference on Computer Networks : The International Journal of Computer and Telecommunications Netowrking*, pages 257–276, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co.

Chakrabarti, S., Punera, K., and Subramanyam, M. (2002). Accelerated focused crawling through online relevance feedback. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 148–159, New York, NY, USA. ACM Press.

Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of the 8th World Wide Web Conference (WWW8)*, pages 1623–1640, Toronto, Canada.

Chau, M. and Chen, H. (2003). Comparison of three vertical search spiders. *Computer*, 36(5):56–62.

Chen, Z., Ma, J., Han, X., and Zhang, D. (2008). An effective relevance prediction algorithm based on hierarchical taxonomy for focused crawling. In Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., and Zhou, G., editors, *Information Retrieval Technology*, volume 4993 of *Lecture Notes in Computer Science*, pages 613–619. Springer Berlin Heidelberg.

Cho, J., Garcia-Molina, H., and Page, L. (1998). Efficient crawling through url ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172.

Choi, Y., Kim, K., and Kang, M. (2005). A focused crawling for the web resource discovery using a modified proximal support vector machines. In Gervasi, O., Gavrilova, M., Kumar, V., Lagan, A., Lee, H., Mun, Y., Taniar, D., and Tan, C., editors, *Computational Science and Its Applications ICCSA 2005*, volume 3480 of *Lecture Notes in Computer Science*, pages 186–194. Springer Berlin Heidelberg.

Daneshpajouh, S., Nasiri, M. M., and Ghodsi, M. (2008). A fast community based algorithm for generating web crawler seeds set. In Cordeiro, J., Filipe, J., and Hammoudi, S., editors, *WEBIST (2)*, pages 98–105. INSTICC Press.

Davison, B. D. (2000). Topical locality in the web. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–279, New York, NY, USA. ACM Press.

Diligenti, M., Coetzee, F., Lawrence, S., Giles, C. L., and Gori, M. (2000). Focused crawling using context graphs. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 527–534, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dumais, S. and Chen, H. (2000). Hierarchical classification of web content. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 256–263, New York, NY, USA. ACM.

Ehrig, M. and Maedche, A. (2003). Ontology-focused crawling of web documents. In *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*, pages 1174–1178, New York, NY, USA. ACM Press.

Gasparetti, F. and Micarelli, A. (2003). Adaptive web search based on a colony of cooperative distributed agents. In Klusch, M., Ossowski, S., Omicini, A., and Laamanen, H., editors, *Cooperative Information Agents*, volume 2782, pages 168–183. Springer-Verlag.

Gasparetti, F., Micarelli, A., and Sansonetti, G. (2014). Exploiting web browsing activities for user needs identification. In *International Conference on Computational Science and Computational Intelligence (CSCI 2014)*. IEEE Computer Society Conference Publishing Services.

Gentili, G., Micarelli, A., and Sciarrone, F. (2003). Infoweb: An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, 17(8-9):715–744.

Hao, H.-W., Mu, C.-X., Yin, X.-C., Li, S., and Wang, Z.-B. (2011). An improved topic relevance algorithm for focused crawling. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 850–855.

Hersovicia, M., Jacovia, M., Maareka, Y. S., Pellegb, D., Shtalhaima, M., and Ura, S. (1998). The shark-search algorithm an application: tailored web site mapping. In *Proceedings of the 7th World Wide Web Conference(WWW7)*, Brisbane, Australia.

Jansen, B. J., Booth, D. L., and Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th annual ACM-SIAM symposium on Discrete algorithms*, pages 668–677, San Francisco, CA, USA.

Liakos, P. and Ntoulas, A. (2012). Topic-sensitive hidden-web crawling. In *Proceedings of the 13th International Conference on Web Information Systems Engineering*, WISE'12, pages 538–551, Berlin, Heidelberg. Springer-Verlag.

Limongelli, C., Sciarrone, F., and Vaste, G. (2011). Personalized e-learning in moodle: The moodle-ls system. *Journal of E-Learning and Knowledge Society*, 7(1):49–58.

Luong, H. P., Gauch, S., and Wang, Q. (2009). Ontology-based focused crawling. In *Information, Process, and Knowledge Management, 2009. eKNOW '09. International Conference on*, pages 123–128.

Menczer, F. and Monge, A. E. (1999). Scalable web search by adaptive online agents: An infospiders case study. In Klusch, M., editor, *Intelligent Information Agents*, pages 323–340. Springer-Verlag, Berlin, Germany.

Menczer, F., Pant, G., and Srinivasan, P. (2004). Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Internet Technol.*, 4(4):378–419.

Menczer, F., Pant, G., Srinivasan, P., and Ruiz, M. E. (2001). Evaluating topic-driven web crawlers. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 241–249, New York, NY, USA. ACM.

Micarelli, A. and Gasparetti, F. (2007). Adaptive focused crawling. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 231–262. Springer Berlin Heidelberg.

Pant, G. and Srinivasan, P. (2005). Learning to crawl: Comparing classification schemes. *ACM Trans. Inf. Syst.*, 23(4):430–462.

Radinsky, K. and Bennett, P. N. (2013). Predicting content change on the web. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 415–424, New York, NY, USA. ACM.

Raghavan, S. and Garcia-Molina, H. (2001). Crawling the hidden web. In *VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases*, pages 129–138, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rungsawang, A. and Angkawattanawit, N. (2005). Learnable topic-specific web crawler. *J. Netw. Comput. Appl.*, 28(2):97–114.

Sakai, T. (2012). Evaluation with informational and navigational intents. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 499–508, New York, NY, USA. ACM.

Srinivasan, P., Menczer, F., and Pant, G. (2005). A general evaluation framework for topical crawlers. *Information Retrieval*, 8(3):417–447.

Zheng, Q., Wu, Z., Cheng, X., Jiang, L., and Liu, J. (2013). Learning to crawl deep web. *Inf. Syst.*, 38(6):801–819.