Adaptive focused crawling using online learning
A study on content related to Islamic extremism

Christos Iliou, Theodora Tsikrika, George Kalpakis, Stefanos Vrochidis, and Ioannis Kompatsiaris

Information Technologies Institute, CERTH
Thessaloniki, Greece
{iliouchristos, kalpakis, theodora.tsikrika, stefanos, ikom}@iti.gr

Abstract. Focused crawlers aim to automatically discover online content resources relevant to a domain of interest by automatically navigating through the Web link structure and selecting which hyperlinks to follow based on an estimation of their relevance to the topic of interest; to this end, classifier-guided approaches are typically employed for identifying hyperlinks having the higher likelihood of leading to relevant content. However, the training data used for building these classifiers might not be entirely representative of the domain of interest, or the domain of interest might change over time. To meet these challenges, this work proposes a novel adaptive focused crawling framework that allows the classifiers that underlie the hyperlink selection policy to be adapted based on the evidence they encounter during their crawls. Our framework uses two different approaches to retrain its models: (i) *Interactive Adaptation*, where a user manually evaluates the discovered resources, and (ii) *Automatic Adaptation*, where the framework uses the already trained classifiers to assess the relevance of newly discovered resources. The evaluation experiments in the domain of Islamic extremism indicate the effectiveness of online learning in focused crawling.

Keywords: Focused crawling, Adaptive learning, Online learning, Islamic extremism

1 Introduction

The proliferation of the use of the Web and social media has greatly facilitated the open diffusion of knowledge and the uninterrupted communication of thoughts, ideas, and interests, thus resulting in a significant increase in the information being shared globally. These new opportunities have also proven very useful for terrorists and extremists for advertising their subversive intentions to broader audiences, cutting across different nationalities, cultures, religions, and residences. Several terrorist organisations and extremist groups have exploited the popularity and broad reach of the Web and social media for supporting their goals of dispersing their propaganda, orchestrating action plans, recruiting new members, and disseminating material targeting potential perpetrators of future attacks.

In this context, the challenge for governments and Law Enforcement Agencies (LEAs) is to better understand and counter potential threats by discovering information related to terrorist and extremist activity on the Internet. To this end, sophisticated Web search and discovery tools are typically employed, such as Web crawlers, which are capable of systematically traversing the Web for the efficient discovery and collection of online content. Starting from a set of predefined (seed) Web pages, Web crawlers fetch (i.e., download) these pages, parse their content for extracting the hyperlinks they contain, and place the extracted hyperlinks in a queue containing the Web pages to be fetched at later stages of the crawl. This process is iteratively repeated until a termination criterion is applied (e.g., a desired number of pages are fetched).

This work aims at presenting a crawler able to discover Web resources on any given topic, with particular focus on topics of interest to LEAs. To this end, we develop a crawler capable of gathering content focused on a given topic by employing a classifier-guided hyperlink selection strategy based on supervised machine learning techniques, that aims to identify the hyperlinks that most likely lead to relevant content. Given the inherently volatile nature of the Web which forms a continuously evolving environment necessitating the employment of dynamically (re)trained classifiers for supporting the hyperlink selection policy and the potential sparsity of appropriate training data for building these classifiers, the focused crawling approach can benefit by incorporating a mechanism allowing for the automatic adjustment of the employed classifiers to new evidence encountered during its crawls. This way, we can ensure that the focused crawler will take into consideration and adapt to any new knowledge emerging about any given topic.

To meet these challenges, we propose a novel adaptive focused crawling framework using online learning that (i) uses a classifier-guided approach for identifying (during the crawling process) hyperlinks having the higher likelihood of leading to relevant content, and (ii) allows these classifiers that underlie the hyperlink selection policy to be adapted based on the evidence they encounter during their crawls. This allows to address the sparsity of appropriate training data, as well as the need to consider the constantly changing landscape. This adaptive focused crawler is able to retrain its hyperlink selection classifiers online either in an automatic or in an interactive manner based on the users' feedback; this has the potential to significantly increase the effectiveness of the static classifiers previously employed for hyperlink selection. The main contribution of our work is thus the development of a generic adaptive focused crawling framework using online learning, which is configured and demonstrated for a specific topic of interest to LEAs: the discovery of Web resources related to Islamic extremism. Our experimental results indicate the significant potential of the proposed approaches on the targeted domain both for the interactive and automatic adaptation settings; the proposed framework thus forms a solid baseline upon which we can further build adaptive focused crawling methods.

The remainder of this paper is structured as follows: Section 2 discusses related work. Section 3 describes the proposed adaptive focused crawling approach.

Section 4 presents the results of the evaluation experiments. Finally, Section 5 concludes this work.

2 Related work

This section first presents state-of-the-art approaches for focused crawling and its application to the terrorism and extremism domain. Subsequently, it discusses the benefits of employing adaptation techniques to the hyperlink selection algorithms of the focused crawling process.

Focused crawlers are used for the automatic discovery of online resources related to a domain of interest by automatically navigating through the Web link structure and selecting the hyperlinks to follow by estimating their relevance to the topic of interest. Focused crawling techniques have been researched for many years [4] and take advantage of the 'topical locality' phenomenon, i.e., that most Web pages link to other pages with similar content [3]. State-of-the-art focused crawlers use classifier-guided approaches based on supervised machine learning techniques for the identification of hyperlinks that have higher likelihood of leading to relevant content [10]. More specifically, a feature vector is calculated for each training sample by typically considering as features the most important terms included in the textual content of the Web pages encountered during the crawls and by estimating the importance of these terms using the tf-idf metric. Each training sample may consist of the anchor text related to the hyperlink of interest, the full content of the parent Web page (i.e., the Web page containing the hyperlink), the text appearing in the vicinity of the hyperlink in the parent Web page (referred to as "surrounding text") or any combination of the above [12]. To this end, state-of-the art approaches typically combine (i) the anchor text of the hyperlink of interest, (ii) a text window of x (e.g., $x = 50$) characters surrounding the anchor text that does not overlap with the anchor text of adjacent hyperlinks, and (iii) the terms extracted from the URL [7, 12].

Focused crawling is a common approach employed for the identification of potentially suspicious online Web content related to the domain of terrorism and extremism. The Dark Web project at the University of Arizona created a suite for Web mining that performs link and content analysis on the Web [1]. Furthermore, researchers have also mentioned that such content is often published on forums, where traditional crawling techniques might fail [5]. To this end, they propose systems that are able to crawl such content from forums [5] and authenticate themselves when needed, provided that a valid username-password combination are provided [7]. To further improve the understanding of terrorist activities, researchers have also focused on collecting and analysing information of Jihad Web sites and develop visualization of their site content, relationships, and activity levels [2]. Furthermore, hybrid focused crawlers have been proposed that are able to traverse both the Surface Web and some of the most popular darknets of the Dark Web (such as Tor); these have been deployed and evaluated for the collection of homemade explosive recipes [7].

All the aforementioned approaches employ models that are trained "statically" based on a fixed set of training samples which cannot be adapted over time. However, two main challenges arise when using pre-trained "static" focused crawlers: (i) the availability of sufficient and appropriate training data for representing the domain of interest and (ii) changes in the domain of interest over time. The latter encompasses the notions of feature-evolution, concept-evolution, and concept-drift [9], where the initially trained classification models may become obsolete over time. To meet these challenges, research has proposed the use of focused crawlers that automatically adapt to the domain of interest.

One of the proposed techniques in the literature, which considers that initially the knowledge about the environment is incomplete, is the use of learning automata, an adaptive decision-making unit [14]. Focused crawlers that use learning automation learn how to choose the optimal actions from a finite set of allowed actions through repeated interactions with a random environment. Such focused crawlers adapt their behaviour based on the feedback they receive (i.e., pages gathered) from the environment. Other approaches perform adaptation by extracting ontologies through an unsupervised adaptive way [6] using a cyclic "maintenance scheme", triggered based on changes in the input data stream, to constantly update their models.

Given the performance gains acquired by employing focused crawling approaches which adapt over time, we propose an adaptive classifier-based focused crawler that shares this advantage. More specifically, this work proposes a novel adaptive focused crawling framework using online learning in two different modes: (i) the *Interactive Adaptation* mode, which takes as input the end user feedback to identify whether a newly discovered Web page was correctly identified as relevant or irrelevant to the domain of interest, and (ii) the *Automatic Adaptation* mode which uses the already trained classifiers for assessing the relevance of newly discovered pages. These newly discovered Web pages are used as input for updating the classification models. The main contribution of this paper is the proposal of an effective technique based on online learning which exploits the advantages of online adaptation and applies it to a classifier-guided focused crawler.

3 Adaptive Focused crawler

This work proposes a novel adaptive classifier-guided focused crawling approach for the discovery of Web resources about a given topic that estimates the relevance of hyperlinks to unvisited resources. The adaptive focused crawler uses state-of-the-art focused crawling approaches and goes one step further by introducing the adaptiveness in the hyperlink selection process based on the evidence encountered during crawling.

The general steps of focused crawling are depicted in Figure 1. Initially, the seed entry points (i.e., the list of Web resources to be crawled) are added to the "frontier", i.e., the list of URLs of the resources not visited yet. In each iteration, a URL is picked from the list and the Web resource corresponding to this

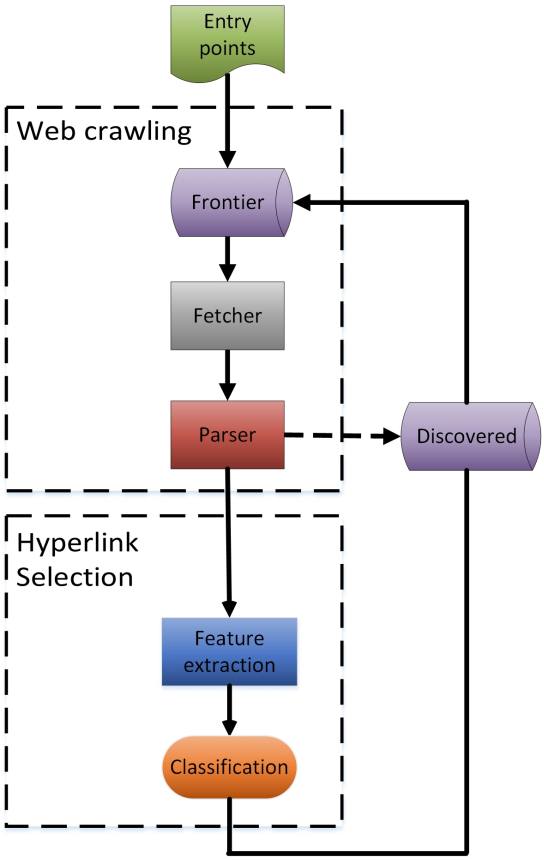

Fig. 1: Classifier-guided focused crawling.

URL is fetched (i.e., downloaded), scraped, and parsed to extract its hyperlinks. Then, the focused crawler estimates the relevance of each hyperlink pointing to an unvisited resource based on the hyperlink's local and global context within its parent Web page. The decision whether a new unvisited hyperlink from a fetched Web resource is relevant to the domain of interest relies on supervised machine learning techniques which train models based on annotated positive and negative samples of online Web content on the domain of interest. Based on these trained classification models, the hyperlink selection process is as follows: each new hyperlink is assigned a confidence score and if this is greater than a threshold, the resource will be considered as relevant and will also be added to the frontier.

Given though that (i) the data to be used for training the classifiers might not be representative of the entire domain, and (ii) changes might occur to the domain of interest over time (e.g., in our case, this could be the introduction of a new terrorist organisation) which might not be covered by the current classifiers,

we propose to adapt the classifiers by training them online using the newly discovered data. Next, the hyperlink selection process (Section 3.1) and the online learning approach (Section 3.2) are described in more detail.

3.1 Hyperlink selection

The hyperlink selection is performed by estimating the relevance of a hyperlink pointing to an unvisited resource with content relevant to the domain of interest. Research has indicated that the combination of the local context of the hyperlink and the global (i.e., full) text of the parent Web resource can effectively predict the relevance of a hyperlink during the selection process [7]. To this end, the current implementation of the hyperlink selection process is performed in a hybrid mode by combining the outcome of a link-based classifier that classifies the hyperlinks based on their local context (i.e., the anchor and surrounding text) and a Web page classifier that classifies the hyperlinks based on their global context (i.e., the entire content of their parent page). This hybrid approach is presented in Figure 2.

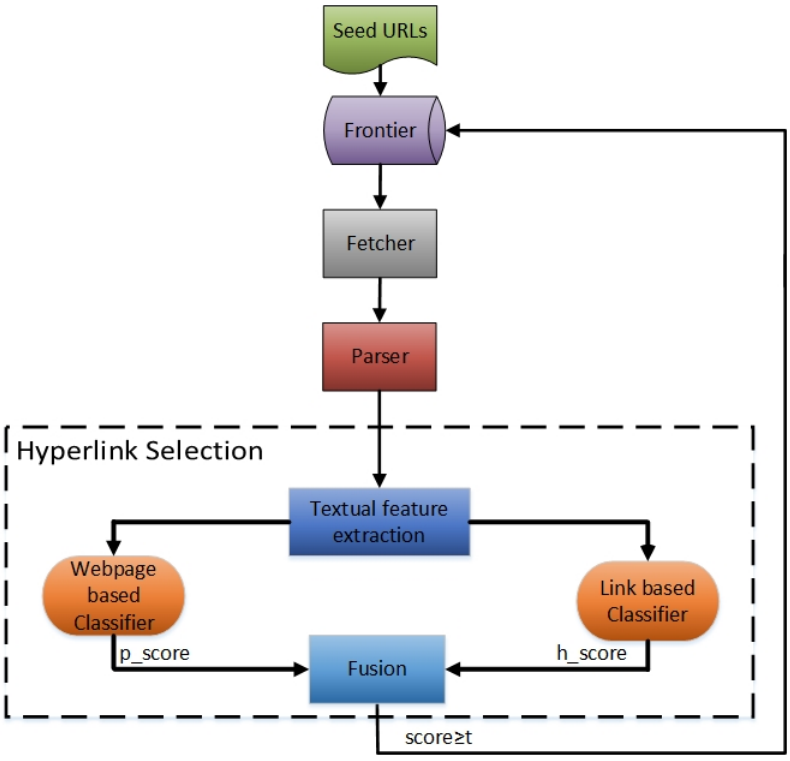

Fig. 2: Hybrid classification architecture.

The first step in the hyperlink selection process is to define a set of features which will most efficiently and meaningfully represent the information that is important for analysis and classification. To this end, our feature extraction relies on textual data. More specifically, each feature corresponds to textual structural components (e.g., words, stems, etc.) found in the context of each resource following a series of processing steps. In particular, given a training set (i.e., a set of Web resources annotated with respect to their relevance to the domain of interest), a lexicon of all terms (features) present in the training set is created by performing the following steps: tokenisation, elimination of stopwords (i.e., non-informative terms that do not facilitate the distinction between dissimilar resoures), and stemming, i.e., reducing the extracted terms to their root form (e.g., "islam", "islamic", "islamism", and "islamist" should all be reduced to the same root form). Once the lexicon is created, each resource can be represented in this feature (vector) space by estimating the term frequency-inverse document frequency (tf-idf) of each of its terms.

Generally, hyperlink classification is a two-step process consisting of a *Training* step, where the classification model is build using the aforementioned training data, and a *Prediction* step, where the generated classification model can be used for assigning class labels to data items in a test dataset. In this case, the class labels indicate the relevance of the hyperlinks to the domain of interest. The effectiveness and accuracy of the classification model is determined by comparing the true class labels in the testing set with those assigned by the model.

A plethora of machine learning approaches have been used for text classification [8]. The selection of the algorithm to be used depends on the type of the problem to be addressed and the available data [8]. In this work, we adopt Support Vector Machines (SVMs) as our classification methodology given their demonstrated effectiveness int the context of both focused crawling applications [11] and document classification [13].

The current implementation of the hyperlink selection process uses the proposed SVM classifiers in two different ways (Figure 2): (i) as a link-based classifier for classifying the hyperlinks based on their local context (i.e., anchor and surrounding text), and (ii) as a Web page classifier for classifying the hyperlinks based on their global context (i.e., the entire content of their parent page). The outputs of these two classifiers are then combined into a single score as follows:

$$score = w_1 \cdot p\_score + w_2 \cdot h\_score \tag{1}$$

where $w_1+w_2 = 1$, p_score is the score produced by the Web page based classifier, and h_score is the score produced by the link-based classifier.

3.2 Adaptive focused crawling

To address (i) the potential initial unavailability of sufficient and appropriate training data for representing the domain of interest, and (ii) the changes occurring in the domain of interest over time, this work proposes a novel adaptive focused crawling approach, where classifiers are retrained periodically by incorporating the newly found samples. The relevance of the new discovered samples

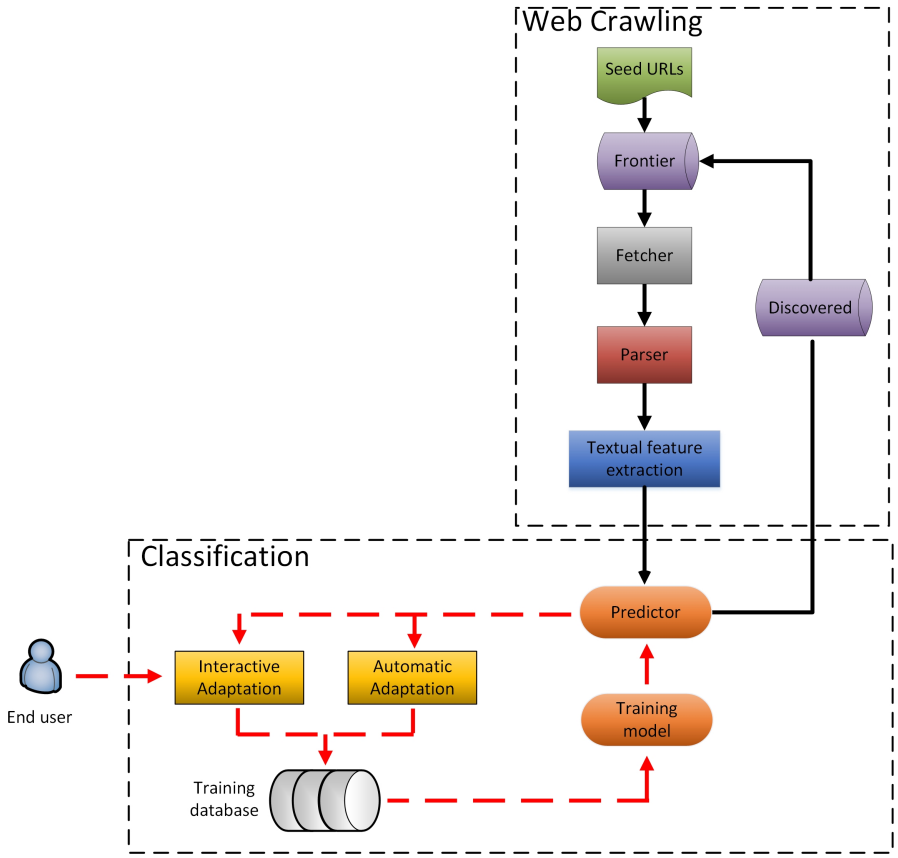

Fig. 3: Adaptive focused crawling.

can be assessed either automatically or by enabling the direct user feedback. The retraining frequency of the framework is based on the number of available data for training. If the newly discovered and annotated data surpass a threshold, the retraining process is initiated. When choosing the threshold value, we should consider that the retraining based only on a small number of newly discovered Web pages will have a minimal effect on the models; on the other hand, retraining the classifiers very rarely would reduce the adaptiveness of the focused crawler to any domain changes. Therefore, the adaptive strategy should keep a balance between performing the retraining process rarely and too frequently, without having a sufficient number of new samples.

To achieve the adaptiveness in our framework, we consider two alternatives of online learning (see Figure 3):

– *Interactive Adaptation* which receives as input the end users' feedback so as to determine whether a newly discovered Web page was correctly identified

as relevant or irrelevant to the domain of interest; these data are used as future training data for adapting the classification model (Algorithm 1).

Algorithm 1 Interactive Adaptation Algorithm

```
1: Let P{} be the set of relevant Web pages in the initial training set
2: Let N{} be the set of non-relevant Web pages in the initial training set
3: Let k be the number of samples needed for the retraining to start
4: j ← 0
5: for each new_web_page do
6:     annotation ← user_feedback(new_web_page)
7:     if annotation == positive then
8:         insert(P, new_web_page)
9:     else
10:        insert(N, new_web_page)
11:    j ← j + 1
12:    if j = k then
13:        retrain_classifiers(P, N)
14:        j ← 0
```

– *Automatic Adaptation* which uses the already trained classifiers to assess the relevance of newly discovered pages; when the classifier identifies Web pages whose classification score is either too high or too low (i.e., the existing evidence indicates the Web page relevance or non-relevance to the domain is strong), it uses this page as future training data for adapting the classification model (Algorithm 2).

Algorithm 2 Automatic Adaptation Algorithm

```
1: Let P{} be the set of relevant Web pages in the initial training set
2: Let N{} be the set of non-relevant Web pages in the initial training set
3: Let k be the number of samples needed for the retraining to start
4: Let t1, t2 be the threshold values strongly indicating relevance, non-relevance
5: j ← 0
6: for each new_web_page do
7:     score ← classification_probability(new_web_page)
8:     if score ≥ t1 then
9:         insert(P, new_web_page)
10:    else if score ≤ t2 then
11:        insert(N, new_web_page)
12:    j ← j + 1
13:    if j = k then
14:        retrain_classifiers(P, N)
15:        j ← 0
```

4 Evaluation

To assess the effectiveness of the proposed adaptive focused crawling approach, a series of experiments were performed in the domain of Islamic extremism. An initial investigation of online content related to Islamic extremism has shown that this type of content is very diverse. Therefore, when employing focused crawlers in this particular case, the data used for training the classifiers might not be sufficiently representative of all the characteristics of the pages of interest. Thus, models need to be retrained periodically, incorporating new pages that are identified as relevant or non-relevant either in an interactive or in an automatic manner. This section describes the experimental setup (Section 4.1) the employed performance metrics (Section 4.2), and the results of the experiments (Section 4.3).

4.1 Experimental setup

A set of 15 seed URLs was used for the experiments; these seed URLs have been provided as Web entry points by experts on the terrorism domain. The content of the selected URLs is relevant to Islamic extremism and the focused crawling experiments that employ them as their seed set aim to discover additional online content relevant to Islamic extremism. The seed set was further split into two subsets, the training set, consisting of 11 URLs, and the testing set consisting of the remaining 4 URLs[1].

Starting from these seeds, a crawl at depth = 1 (where depth is the maximum distance allowed between seed and crawled pages) was performed and the discovered Web pages, along with their content, were stored locally. The total number of Web pages collected during the crawl was 179 Web pages: 138 Web pages were discovered from the training seed set and 41 from the testing seed set. The collected URLs were annotated based on their relevance to the domain of Islamic extremism; this resulted in having a training set with 80 positive and 58 negative samples, and a test set with 23 positive and 18 negative samples.

Starting from the four seeds in the test set, the hyperlink selection process estimated the relevance of the encountered hyperlinks by combining two classifiers: (i) a link-based classifier that takes advantage of the local context (i.e., anchor text and surrounding text) of the hyperlinks in the parent page, and (ii) a Web page classifier that exploits the content of the parent Web page; see Section 3.1 for further details. The total score was calculated using a combination of the scores produced by the Web page and the link-based classifiers, with $w_1 = 0.1$ and $w_2 = 0.9$, respectively (see Section 3.1). If the estimated score was greater than $t = 0.5$, the hyperlinks were considered as relevant.

The effectiveness of the proposed adaptive focused crawling approach was evaluated in three stages. The first stage represents the initial point of the focused crawl where the classifier employed is built using a training set consisting of the

[1] The actual URLs are not provided so as to avoid the inclusion of potentially sensitive information, but can be made available upon request.

80% of the total training samples (i.e., the initial classifier is already built before the beginning of the crawl process). The second and the third stage represent the points where the adaptive online training takes place (i.e., at some point during the focused crawl after discovering new pages). In these two stages, an additional 10% of the remaining samples on the original training dataset, respectively, is employed for retraining the initial classifier; the new samples provided simulate the new Web pages discovered during the course of the crawl. Our experiments simulate an online training process where the data are periodically imported to the training set after $k = 14$ new pages are encountered (i.e., 14 additional Web pages were used for retraining the classifiers at each of the second and the third stages, respectively). Furthermore, the classifiers from each stage were tested on the same data. The number of samples employed for training at each stage of our experiments and the number of samples used for testing is illustrated in Table 1.

Table 1: Number of positive and negative samples in the training and test datasets.

	Training data			**Test data**
	80%	90%	100%	
Positive	64	72	80	23
Negative	46	52	58	18
Total	110	124	138	41

We tested the two online learning approaches implemented in our work (i.e., the Interactive Adaptation and the Automatic Adaptation). In both cases, the first stage of our experiments included the training of the initial classification models (the link-based and Web page classifiers, respectively) on 80% of the available training data. During the second and third stages, the training set was further enriched with an additional 10% of the remaining training data, respectively (i.e., 90% and 100% of the total training samples were used for the second and third stage, respectively).

In the Interactive Adaptation mode, our goal was to simulate the end user feedback provided during the second and third stage of our experiments; therefore, the adaptive models were produced based on the annotation performed by domain experts for all the new pages discovered during the second and the third stages. On the other hand, given that the Automatic Adaptation mode is based on the already trained classifiers for evaluating the relevance of newly discovered pages, it performs the retraining during the second and the third stage by adding new entries for which the estimated relevance or non-relevance to the domain of Islamic extremism is strong enough. Specifically, the newly discovered samples are used as positive samples in the retraining only when their estimated score is greater that a value t_1, or as negatives when their score is lower than a value t_2, indicating strong confidence of the classifiers; these values were set as $t_1 = 0.8$ and $t_2 = 0.2$.

Finally, the classifiers were implemented using the LIBSVM package of the Weka machine learning software[2], with the text-based classification scores corresponding to its probabilistic outputs; their parameters are presented in Table 2.

Table 2: Parameters used for the SVM classifiers.

SVM parameters	
SVM type	C-SVM
Kernel	Radical basis function
Penalty parameter	1
Gamma	1/n, where n is the number of features
Tolerance	0.001
Shrinking heuristics	enabled

4.2 Evaluation metrics

To assess the effectiveness of our approach we calculated the precision and recall metrics as well as their harmonic mean, F-measure. Precision is the fraction of relevant pages among all the fetched pages, while recall is the fraction of relevant pages fetched over the total amount of relevant pages. The latter requires knowledge of all relevant pages on a given topic which is a practically infeasible task in the context of the Web. To address this limitation, two recall-oriented evaluation techniques have been proposed [10]: (i) manually designate a few representative pages on the topic and measure the fraction of them discovered by the crawler and (ii) measure the overlap among independent crawls initiated from different seeds to see whether they converge on the same set of pages; only the former could be applied to our experimental set-up.

4.3 Results

Table 3 presents the results for the Interactive Adaptation for the three evaluation stages, when the 80%, 90% and 100% of the training set was used, respectively. As the number of the training data increases, the overall precision also increases, reaching over 90% when all available training samples are used (i.e., in the third stage). This is accompanied by a small decrease in recall, indicating that a few relevant Web pages are missed in the effort to reach a high precision. Nevertheless, the overall performance in terms of the F-measure also improves.

Contrary to the Interactive Adaptation, the Automatic Adaptation does not use all the discovered data for training but only the data whose classification score entails strong confidence (i.e. when the score is either too high or too low). Specifically, we only use as positive samples, data with classification scores greater than 0.8, and as negative samples, data with classification scores less than

[2] http://www.cs.waikato.ac.nz/ml/weka/

Table 3: Evaluation results for Interactive Adaptation of the hyperlink selection classifiers.

	Data for online training		
	80%	90%	100%
# Positive	64	72 (+8)	80 (+8)
# Negative	46	52 (+6)	58 (+6)
# Total	110	124 (+14)	138 (+14)
Precision	0.61	0.70	0.91
Recall	1.00	1.00	0.95
F-measure	0.76	0.82	0.93

Table 4: Evaluation results for Automatic Adaptation of the hyperlink selection classifiers.

	Data for online training		
	80%	90%	100%
# Positive	64	68 (+4 out of 8)	69 (+1 out of 8)
# Negative	46	47 (+1 out of 6)	48 (+1 out of 6)
# Total	110	115 (+5 out of 14)	117 (+2 out of 14)
Precision	0.61	0.70	0.78
Recall	1.00	0.94	0.95
F-measure	0.76	0.80	0.86

0.2 (the values between 0.2 and 0.8 are not considered to provide strong evidence, and hence the respective Web pages are not used for retraining purposes). The total amount of data considered as positive and negative samples (out of all the available data) at each retraining stage are shown in Table 4, together with the results of the evaluation experiments.

Similarly to the above, as the number of the training data increases, the overall precision and F-measure also increase, while recall slightly drops. The achieved results indicate the potential benefits of considering automatically identified training samples; however, additional large-scale experiments are needed so as to reliably assess the potential effectiveness of the proposed approaches.

5 Conclusions

This work proposed a novel adaptive focused crawling framework using online learning that (i) uses a classifier-guided approach for identifying (during the crawling process) hyperlinks having the higher likelihood of leading to relevant content, and (ii) allows these classifiers that underlie the hyperlink selection policy to be adapted based on the evidence they encounter during their crawls. This adaptive focused crawler is able to retrain its hyperlink selection classifiers online either in an automatic or in an interactive manner based on user's feedback. Experiments in the domain of Islamic extremism indicate the significant potential of the adopted approaches on the targeted domain, given that our framework has

demonstrated satisfactory performance, both for the interactive and automatic adaptation settings. Our future work will investigate strategies for the selection of the optimal subset of all available Web pages for online training for further increasing our framework's effectiveness and efficiency.

Acknowledgements

This work was supported by the TENSOR project (H2020-700024), funded by the European Commission.

References

1. Chen, H.: Dark web: Exploring and data mining the dark side of the web, vol. 30. Springer Science & Business Media (2011)
2. Chen, H., Chung, W., Qin, J., Reid, E., Sageman, M., Weimann, G.: Uncovering the dark web: A case study of jihad on the web. Journal of the Association for Information Science and Technology **59**(8), 1347–1359 (2008)
3. Davison, B.D.: Topical locality in the web. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 272–279. ACM (2000)
4. De Bra, P.M., Post, R.: Information retrieval in the world-wide web: Making client-based searching feasible. Computer Networks and ISDN Systems **27**(2), 183–192 (1994)
5. Fu, T., Abbasi, A., Chen, H.: A focused crawler for dark web forums. Journal of the Association for Information Science and Technology **61**(6), 1213–1231 (2010)
6. Hassan, T., Cruz, C., Bertaux, A.: Ontology-based approach for unsupervised and adaptive focused crawling. In: Proceedings of The International Workshop on Semantic Big Data. p. 2. ACM (2017)
7. Iliou, C., Kalpakis, G., Tsikrika, T., Vrochidis, S., Kompatsiaris, I.: Hybrid focused crawling for homemade explosives discovery on surface and dark web. In: Availability, Reliability and Security (ARES), 2016 11th International Conference on. pp. 229–234. IEEE (2016)
8. Khan, A., Baharudin, B., Lee, L.H., Khan, K.: A review of machine learning algorithms for text-documents classification. Journal of advances in information technology **1**(1), 4–20 (2010)
9. Masud, M.M., Chen, Q., Khan, L., Aggarwal, C., Gao, J., Han, J., Thuraisingham, B.: Addressing concept-evolution in concept-drifting data streams. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. pp. 929–934. IEEE (2010)
10. Olston, C., Najork, M.: Web crawling. Foundations and Trends in Information Retrieval **4**(3), 175–246 (2010)
11. Pant, G., Srinivasan, P.: Learning to crawl: Comparing classification schemes. ACM Transactions on Information Systems (TOIS) **23**(4), 430–462 (2005)
12. Pant, G., Srinivasan, P.: Link contexts in classifier-guided topical crawlers. IEEE Transactions on knowledge and data engineering **18**(1), 107–122 (2006)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34**(1), 1–47 (2002)
14. Torkestani, J.A.: An adaptive focused web crawling algorithm based on learning automata. Applied Intelligence **37**(4), 586–601 (2012)