Iliou et al. EURASIP Journal on Information Security (2017) 2017:11
DOI 10.1186/s13635-017-0064-5

EURASIP Journal on Information Security

RESEARCH Open Access

Hybrid focused crawling on the Surface and the Dark Web

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

Abstract

Focused crawlers enable the automatic discovery of Web resources about a given topic by automatically navigating through the Web link structure and selecting the hyperlinks to follow by estimating their relevance to the topic of interest. This work proposes a generic focused crawling framework for discovering resources on any given topic that reside on the Surface or the Dark Web. The proposed crawler is able to seamlessly navigate through the Surface Web and several darknets present in the Dark Web (i.e., Tor, I2P, and Freenet) during a single crawl by automatically adapting its crawling behavior and its classifier-guided hyperlink selection strategy based on the destination network type and the strength of the local evidence present in the vicinity of a hyperlink. It investigates 11 hyperlink selection methods, among which a novel strategy proposed based on the dynamic linear combination of a link-based and a parent Web page classifier. This hybrid focused crawler is demonstrated for the discovery of Web resources containing recipes for producing homemade explosives. The evaluation experiments indicate the effectiveness of the proposed focused crawler both for the Surface and the Dark Web.

Keywords: Focused crawling, Dark web, Darknets, Tor, I2P, Freenet, Dynamic linear combination

1 Introduction

The terror veil spread over Europe the past years has put increasing pressure to government authorities, law enforcement agencies, and intelligence services, so as to uphold the rule of law and keep the citizen safe by raising their awareness against local and international terrorist and organized crime groups to the highest level possible. The repeated terrorist attacks targeting mainly crowded areas visited by civilians for their everyday commute, work, or entertainment causing a hecatomb of innocent victims indicate clearly that the war against terrorism should be fought with all the available means, from enforcing stricter policies in border control in an effort to identify possible terrorist perpetrators to performing intelligence gathering, not only by taking advantage of the conventional methodologies applied in threat assessment, prevention, and mitigation but also by considering alternative solutions to meet these demands. The continuous technological advancement provides a plethora of additional opportunities for law enforcement and security agencies to tackle the contemporary threats and challenges in a more effective and efficient manner.

* Correspondence: iliouchristos@iti.gr
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece

At the same time, the rapid penetration of broadband services in the average household over the past 10 years, along with the abundance of online resources for storing content, has resulted in an enormous increase in the use of the Internet, and in the proliferation of the information being shared globally. This growth has facilitated the communication and the diffusion of knowledge and experience among users worldwide in many different domains and disciplines of vital significance for the humanity; however, it has also proven very useful for sharing information online in ways which pose a threat for the society. Extremist groups and terrorist organizations exploit the numerous additional opportunities provided by the Internet in their effort not only to spread their propaganda, radicalize new members, or organize their strategic operations but also to disseminate content with potentially subversive use, including information for manufacturing homemade explosives (HMEs) and improvised explosive devices (IEDs), for supporting small extremist cells and lone-wolf terrorists preparing and committing acts of violence and terrorism outside the general command structure of an organization.

In this context, Law Enforcement Agencies (LEAs) have emphasized discovering information related to

terrorist activity on the Internet by taking advantage of the most recent advancements in the field of Web search. Their efforts have focused on the so-called Surface Web representing the part of the Web gathered and indexed by conventional general-purpose search engines, such as Google, Yahoo!, or Bing. However, such search engines index just a small portion of the available Web information; the rest non-indexable content lies in the so-called Deep Web which in general includes content that cannot be retrieved by the Web crawlers employed by the conventional search engines due to several limitations (such as dynamic content generated in response to queries and private content requiring authorized access). There is also a part of the Deep Web, known as the Dark Web, which is intentionally hidden and is virtually inaccessible through typical Web browsers; it can only be reached if the appropriate setup of software, configuration, or authorization is used [1]. The Dark Web provides anonymity both from a user and a data perspective and thus has become a popular solution not only for the communication of secret and sensitive data for legitimate reasons but also for sharing illegal material and disseminating extremist content. It is formed by several *darknets* (e.g., Tor[1], I2P[2], and Freenet[3]), providing encrypted, decentralized, and anonymous network communication. Due to the technological restrictions required for entering the world of these darknets, the crawling techniques applied in the Dark Web are more challenging than the ones applied in the Surface Web.

As a result of the distinctive nature of the Dark Web as well as of the volatility of the Web sites hosted (i.e., typically, the Dark Web content is hosted on machines not preserving a 24/7 uptime), the conventional search engines do not index the content of these darknets; there is just a small number of stable search tools for indexing part of the Dark Web, (i.e., search engines targeting content on specific darknets, such as Ahmia[4] or Torch[5]), since the ephemeral nature of most Dark Web content does not allow for providing a reliable search infrastructure. The most prominent way for finding Dark Web entry points is by taking advantage of directory listings, known as hidden wikis or hidden service lists, which advertise URLs of Web sites hosted on several darknets of the Dark Web, categorized based on their relevance to thematic domains, mainly related to illegal activities (e.g., HMEs, weapons, terrorism, child pornography, human trafficking, and drugs). These listings can be found not only on pages hosted on the darknets of interest but also on several repositories hosted on the Surface Web. At the same time, the darknets do not constitute isolated islands, unlinked from the outer world. It is very common for a Web page hosted on a darknet to contain hyperlinks pointing to content hosted either on different darknets or on the Surface Web. The interconnectivity existing among the Surface and the Dark Web, as well as between the different darknets, dictates that the crawling process employed should be able to traverse both the Surface and the Dark Web and seamlessly propagate among the different network types encountered.

This work aims to develop a crawler capable of traversing both the Surface Web and the darknets present in the Dark Web (i.e., Tor, I2P, and Freenet) with the goal of discovering Web resources on any given topic, with particular focus on topics of interest to LEAs. To this end, it develops a crawler capable of gathering content focused on a given topic by selecting to follow only the hyperlinks that lead to relevant resources. This selection is performed using classification approaches that take advantage of the local context of each hyperlink, as well as of the global context of the parent and the destination page of a hyperlink. The proposed focused crawler is demonstrated for a specific topic of interest to LEAs: the discovery of Web resources containing recipes for producing HMEs.

This topic has been investigated in the context of the activities of the HOMER[6] (HOmeMade Explosives and Recipes characterization) FP7 EU project. An empirical study conducted by domain experts and law enforcement agents participating in HOMER has shown that, in addition to the numerous HME Web resources available on the Surface Web, the anonymous nature of the Dark Web facilitates the publishing of HME-related information, and hence, a crawler focused on the HME domain should be capable of traversing both the Surface and the Dark Web. This study also revealed that for the HME domain, there is an interconnection between the Surface and the Dark Web, as well as among the different darknets of the Dark Web. For example, Web sites containing HME-related information hosted on the Surface Web are likely to include hyperlinks to relevant content hosted on different darknets, such as Tor or I2P, and vice versa. This entails that it is really important for an HME-related crawler to be capable not only of traversing separately the several darknets present in the Dark Web but also of dealing with the interconnectivity encountered between the different darknets and/or the Surface Web during a single crawl. Lastly, this study also indicated that the local context of hyperlinks pointing to Tor and/or Freenet Web pages may not contain sufficient HME-related text; therefore, effective link selection should rely not only on the local context of hyperlinks, but also on additional evidence.

In this context, this work proposes a focused crawler capable of following hyperlinks both on the Surface and several darknets of the Dark Web (i.e., Tor, Freenet, I2P) and also between them in a hybrid manner. The

proposed crawler employs a modularized infrastructure responsible for automatically forwarding the traffic destined for each network to the respective fetching module, which is capable of enabling and managing the crawler's communication with the Web resources residing in each separate network. The main contribution of our work is the development of a generic hybrid focused crawling framework, able to traverse seamlessly the Surface Web and several darknets present in the Dark Web during a single crawl by automatically adapting its crawling behavior and its hyperlink selection strategy based on the destination network type and/or the strength of the local evidence present in the vicinity of the hyperlinks visited. The selection of the hyperlinks to follow is guided by three classifiers each one serving a different purpose: (i) a link-based classifier taking into account the local context of the hyperlinks encountered, i.e., the anchor text, the surrounding text, and the terms within the URL of each hyperlink; (ii) a parent Web page classifier estimating the relevance of the parent page of a hyperlink to the domain of interest based on its global textual context; and (iii) a destination Web page classifier producing a relevance score about the page a hyperlink points to using its global textual context. Motivated by the activities of the HOMER project, this framework was configured towards the discovery of resources on the Surface and the Dark Web containing HME recipes.

As an extension of a preliminary study [2], this work has the following additional contributions:

1. It expresses the proposed focused crawling approach employing the three classifiers in an individual or in a combined mode based on conditions related to the destination network of a hyperlink (e.g., the hyperlink selection policy differentiates when a link points to the Dark Web or to specific darknets of the Dark Web) and/or to whether the local context representation of the hyperlinks encountered conveys meaningful or sufficient information.
2. It investigates 11 different hyperlink selection methods utilizing the link-based classifier or a combination of the link-based classifier with the parent or the destination Web page classifier, enabled based on the conditions discussed earlier.
3. It proposes a novel hyperlink selection strategy, which is based on a dynamic linear combination among the link-based classifier and the parent Web page classifier. To the best of our knowledge, this dynamic combination approach proposed is a new, untested strategy, which aims at dynamically adjusting the confidence score generated by the link-based classifier depending on its distance from the parent classifier score, when there is evident discordance between the two scores produced.
4. Finally, it performs larger-scale experiments for assessing the proposed focused crawler performance in terms of its effectiveness and efficiency for a specific topic of interest from the LEA perspective: the discovery of Web resources containing information about HME recipes.

The remainder of this paper is structured as follows: the "Related work" section reviews related work. The "Hybrid focused crawler architecture" section discusses the distinct characteristics of the proposed hybrid focused crawler, giving emphasis on the components where a different approach, compared to that employed by a typical focused crawler, has been followed. The "Frontier component" section describes in detail the hyperlink selection policy employed. Fetcher component section presents the results of the evaluation experiments for the HME domain. Finally, the "Link selection component" section discusses our conclusions.

2 Related work

This section first discusses the state-of-the-art approaches for focused crawling and crawling on the Surface and the Dark Web and then reviews the most important research efforts for discovering terrorist or extremist-related Web content.

Focused (or topical) crawlers allow for the selective discovery of Web resources related to a given topic by automatically traversing the Web graph structure by only following the hyperlinks which are estimated to point to other resources relevant to the topic of interest. The process starts by defining a set of seed Web pages (i.e., the starting points of the crawl) relevant to the topic of interest and adding them to the frontier [3], i.e., the list containing the Web page URLs already discovered but not yet downloaded by the crawler. Each page included in the frontier is fetched (i.e., downloaded) and parsed for extracting the hyperlinks it contains, so that the crawler can select the ones that most plausibly point to other pages relevant to the topic. Each selected hyperlink is subsequently added to the frontier, and this process is iteratively repeated until a termination criterion is satisfied (e.g., a desired number of pages are fetched, or the limit on the crawling depth is reached).

Predicting the benefit of fetching an unvisited Web page is a challenging task; focused crawlers exploit the "topical locality" observation on the Web, which dictates that most Web pages tend to link to other pages with related content. To this end, state-of-the-art approaches [3] adopt classifier-guided crawling strategies based on supervised machine learning methods which rely on two sources of evidence for selecting the hyperlinks to be followed in order to reach relevant Web resources: (i) the *local* context of the hyperlinks, usually represented

by the textual content appearing in their vicinity within their parent page, such as their anchor text and (part of) their surrounding text, and/or (ii) the *global* context of hyperlinks, typically represented by evidence associated with the entire parent page, such as its textual content or its hyperlink structure [4].

At the same time, several crawlers for the Dark Web have been developed as a result of relevant research efforts. A crawler capable of navigating within the Tor network while providing a level of anonymity for hiding its identity has been developed in the context of the Artemis project, in an effort to monitor and examine the most significant properties of the Dark Web [5]. Moreover, within the context of a DARPA project, a savvy crawler for traversing the Tor network by adjusting its behavior based on the reachability of each resource encountered has been built, motivated by the need for extracting and analyzing the content hosted on hidden services [6]. Further research efforts have proposed a focused crawling system navigating in extremist Dark Web forums by employing a human-assisted accessibility approach so as to gain access to the forums of interest [7]. Finally, a Web crawler for Tor has been developed in order to analyze the content and research the popularity of the most prominent content categories on Tor hidden services [8]. Contrary to the above, the focused crawling approach proposed provides additional functionalities apart from traversing the Tor network and is capable of navigating through the Surface Web and additional darknets of the Dark Web, such as I2P and Freenet, by automatically adapting its behavior and link selection policy based on the destination network type of a hyperlink and/or the strength of the local context around a hyperlink.

Regarding the discovery of terrorist-related Web resources, the Dark Web project at the University of Arizona has provided the most comprehensive multilingual suite of text and Web mining tools for performing link and content analysis aiming at studying and understanding terrorist and extremist phenomena [9]. Furthermore, a methodology for collecting and analyzing Dark Web information has been applied on a set of Jihad Web sites, with the aim to aid the process of intelligence gathering and to improve the understanding of terrorist and extremist activities [10]. Both projects have addressed the whole breadth of terrorist and extremist content, rather than HME information, as done here. In addition, these research efforts have addressed the Dark Web in a different context than the one provided within this work; with the term "Dark Web," they refer to the part of the Surface Web helping to achieve the subversive objectives of terrorists and extremists by publishing relevant content on various forms, including Web sites, forums, blogs, social network sites, and virtual world sites.

With regard to research efforts related to discovering and analyzing HME Web content, a concept detection mechanism has been developed in the context of HOMER project with the goal of identifying the relevance of already discovered multimedia files (videos/images) to the HME domain in an automatic fashion [11]; however, this work has solely addressed the identification of HME-related objects in multimedia, rather than the discovery of such content on the Web. In addition, a Knowledge Management Platform for managing the discovery, analysis, and retrieval of HME-related content has been developed [12]; nevertheless, this effort has mainly addressed issues related to the architecture of the entire framework for the HME knowledge management, rather than the discovery of HME-related Web content. Moreover, a work related to the development of an interactive search engine for the discovery of HME-related information on the Web has been proposed [13]; however, this effort mainly deals with the interaction of the end users with the framework, rather than with the approaches implemented for the discovery of the HME information.

Contrary to the aforementioned studies, this work proposes a hybrid focused crawler seamlessly navigating through both the Surface and the Dark Web, capable of implementing several different link selection methods employed under different conditions related to the destination network type and/or the local context strength of a hyperlink, according to the desirable outcome of the discovery process. The main advantage of our hybrid crawler infrastructure lies on its ability to effectively deal with resources belonging to different darknets of the Dark Web or to the Surface Web during a single crawl. This generic framework is applied here to the HME domain; however, it can easily be configured to deal with any domain of interest.

3 Hybrid focused crawler architecture

The proposed hybrid crawler is capable of traversing both the Surface Web and several darknets of the Dark Web (i.e., Tor, I2P, Freenet) and adapting its crawling behavior based on the network encountered (i.e., Surface or Dark Web). It is based on a hyperlink selection policy capable of supporting several different methods employed under different conditions, based on the destination network type of each hyperlink visited and/or the local context present in the vicinity of a hyperlink. To this end, it exploits a suite containing three different classifiers which are utilized depending on the hyperlink selection policy employed: (i) a link-based classifier for estimating the relevance of a hyperlink to an unvisited resource based on its local context on the parent page,

(ii) a parent Web page classifier which estimates the relevance of the parent page containing a hyperlink based on its global context, and (iii) a destination Web page classifier that estimates the relevance of the destination page a hyperlink points to based on its actual textual content. The three different classifiers are used either individually or in a combined mode where the link-based classifier is accompanied by the parent or the destination Web page classifier when certain conditions are satisfied. The developed focused crawler is based on a customized version of Apache Nutch[7] (version 1.9).

An overview of the proposed crawling approach is depicted in Fig. 1. First, the seed pages are added to the frontier list. In each iteration, a URL is picked from the frontier and is forwarded to the responsible fetching module based on its network type (i.e., Surface Web, Tor, I2P, Freenet). The page corresponding to this URL is fetched and parsed to extract its hyperlinks which are then forwarded to the link selection component. Then, the focused crawler estimates the relevance of each hyperlink pointing to an unvisited page based on the selection method employed. Each available method requires the use of different classifiers (or combinations of them) enabled under certain conditions for succeeding the desirable outcome. Next, the main components of our hybrid focused crawler are presented.

4 Frontier component

The frontier contains the URLs discovered but not yet visited (i.e., downloaded) during the crawl process. It is initialized with a predefined set of seed URLs relevant to the domain of interest (HME recipes in this work), and it is iteratively updated with new unvisited URLs, based on the hyperlink selection method employed. The frontier of our hybrid focused crawler supports all the different types of URLs encountered in the supported darknets of the Dark Web (i.e., Tor, I2P, and Freenet), along with the typical URLs found on the Surface Web.

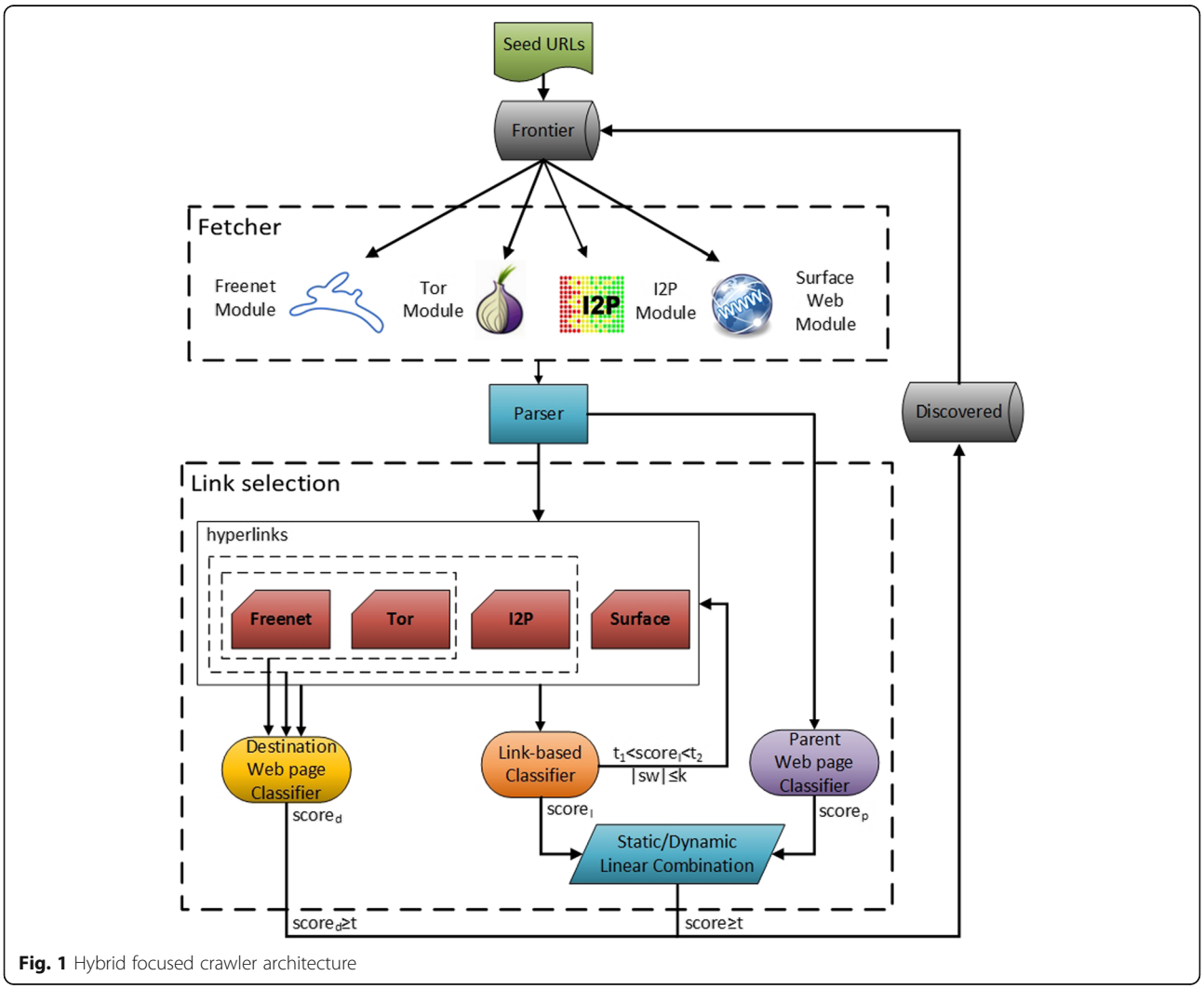

Fig. 1 Hybrid focused crawler architecture

5 Fetcher component

The fetcher component is configured appropriately in order to support fetching Web pages both from the Surface Web and the darknets of the Dark Web.

As a result of the technological restrictions existing when accessing the different darknets of the Dark Web (i.e., Tor, I2P, and Freenet), providing a universal solution capable of successfully fetching the Web pages both from the Surface and the Dark Web is a challenging task; each darknet requires utilizing the appropriate service enabling the communication of the fetcher with the Web resources residing within each network. Specifically, Tor service is required when accessing the Tor hidden services, the I2P service is required for communicating with i2p pages part of Eepsites hosted on I2PTunnel "servers", and the Freenet proxy is needed for fetching freesite pages distributed among one or more Freenet nodes. To this end, the fetcher component consists of four separate fetching modules, each being responsible for handling requests for the Surface Web, the Tor network, the I2P network, and Freenet, respectively, after utilizing the services required.

The URLs encountered during a crawl are automatically distinguished by the fetcher component based on their network type (i.e., Surface Web, Tor, I2P, and Freenet URLs), and the respective traffic is forwarded to the responsible fetcher module, so as to proceed with downloading the content of the Web resource under consideration. Furthermore, the fetcher component employs a human-assisted accessibility approach for gaining access to Web pages belonging to Web site(s) that require authentication based on an auto log-in functionality. A prerequisite for enabling this operation is to acquire a valid username and password for the Web site(s) of interest and insert it into the crawler's configuration file. The auto-login process stores locally the cookie produced from the authentication process for every Web site of interest and re-transmits it every time a new request for a Web page (being part of the specific Web site(s)) is submitted. This infrastructure (see Fig. 1) enables the focused crawler to successfully visit Web pages belonging to the Surface and/or the Dark Web during a single crawl and cope with the possible interconnectivity existing among them.

6 Link selection component

The hybrid focused crawler selects the hyperlinks to follow by employing a classifier-guided approach, which relies on a suite consisting of three different classifiers: (i) a link-based classifier, (ii) a parent Web page classifier, and (iii) a destination Web page classifier. The link-based classifier estimates the relevance of a hyperlink to an unvisited resource based on its local context in the parent page, whereas the parent and the destination Web page classifiers estimate the relevance of the parent and the destination page of a hyperlink, respectively, based on their global (textual) context. These three classifiers are employed by the hyperlink selection methods supported within the link selection component. Each method employs a different combination of the available classifiers, where the link-based classifier is accompanied by the parent or the destination Web page classifier, depending on conditions related to the destination network, the link score produced, or the existence of strong evidence within the local context of a hyperlink.

Next, the hyperlink selection policy is further analyzed focusing on the methods supported.

7 Hyperlink selection policy

As already discussed, the hyperlink selection policy relies on three different classifiers (i.e., link-based, parent Web page and destination Web page classifier), combined according to conditions related to the destination network type and/or the strength of local evidence around a hyperlink on a parent page. Based on recent research [14] and the empirical study conducted in the context of HOMER which indicated that the anchor text and also the URLs of hyperlinks leading to HME information often contain HME-related terms, e.g., the name of the HME, the link-based classifier represents the local context of each hyperlink using: (i) its anchor text, (ii) a text window of x characters ($x = 50$) surrounding the anchor text that does not overlap with the anchor text of adjacent hyperlinks, and (iii) the terms extracted from the URL. Given that Tor and Freenet URLs inherently contain network-specific terms within the URL domain name which do not convey any meaningful information (i.e., onion URLs contain automatically generated 16-character alpha-semi-numeric hashes, whereas Freenet URLs contain a localhost IP address), the local context representation in the case of Tor or Freenet URLs includes the URL terms extracted only from the URL path or parameters. Each hyperlink's local context is represented (after stopwords removal and stemming) using a *tf.df* term weighting scheme, where *tf(t,d)* is the frequency of term t in sample d, normalized by the maximum frequency of any t in that sample, and *df(t)* is the number of samples containing that term in the collection of samples.

The classification of this local context is performed using supervised machine learning based on Support Vector Machines (SVMs), given their demonstrated effectiveness in such applications [15]. The confidence score for each hyperlink (i.e., $score_l$) is obtained by applying the trained classifier on its feature vector. If the score is above a given (experimentally tuned) threshold t, the page is considered relevant to the domain of interest.

However, the hyperlinks' local context may not contain sufficient evidence that will allow the link-based classifier to produce strong estimations about the relevance of a hyperlink to the domain of interest. The HOMER empirical study for the HME domain indicated that this is particularly likely for hyperlinks pointing to Dark Web pages (i.e., Tor, I2P, Freenet), since several hyperlinks which are included in HME-related pages and point to Dark Web pages are part of URL directory lists, which in turn entails that there is no or insufficient textual evidence in the vicinity of the these hyperlinks. Furthermore, as discussed earlier, Tor and Freenet URLs contain network-specific terms which are not included in the hyperlinks' representation, and hence, in several cases, the number of the terms extracted may not be sufficient for representing the hyperlink.

Based on the observation that the local context of a hyperlink may lack strong evidence and thus may affect negatively the link-based classifier's performance, additional link selection methods are investigated in an effort to improve the effectiveness of the hybrid focused crawler. In this context, the link-based classifier is accompanied by a parent or a destination Web page classifier, so as to further enhance the link selection policy. As already discussed, given that Web pages tend to link to others with similar content, the global context of hyperlinks on the parent page could be used for adjusting the estimates of relevance determined by the local context; hence, the parent Web page classifier can be used in conjunction with the link-based classifier in an effort to further enhance the assessment process [16]. On the other hand, a destination Web page classifier can be used for reaching concrete estimations about a hyperlink's destination page relevance to the domain of interest, given that the actual textual context of a page constitutes the most representative means for producing solid estimations. However, this solution entails the potential extra burden of downloading the destination page content; hence, it may deteriorate the focused crawler time performance; thus, it should be sparingly used, only under conditions indicating that the benefit acquired will exceed the efficiency reduction.

The parent and the destination Web page classifiers both take advantage of the textual content present on a hyperlink's parent and destination Web page, respectively, so as to estimate their relevance to the HME domain. Each resource is parsed; its textual content is extracted; tokenization, stopwords removal, and stemming are applied; and a textual feature vector is generated using the *tf.idf* term weighting scheme, where *tf(t,d)* is defined as above and *idf(t)* is the inverse document frequency of term t in the collection. Each classifier produces a confidence score using different input: the parent page and the destination textual content, respectively.

The conditions upon which the parent and the destination Web page classifiers are enabled vary and depend on the tradeoff among the advantages and the disadvantages of each approach. Given that the local context of the hyperlinks pointing to the Dark Web often contains insufficient evidence, the destination Web page classifier may be enabled based on the network type encountered (i.e., Surface Web, Tor, I2P, Freenet) either when the link-based score is not robust enough (i.e., it produces a score which does not provide a strong positive or negative estimation about the relevance of the page with the domain of interest) or when the number of the words contained within the local context is not sufficient for producing a concrete estimation. In the former case, a weak hyperlink estimation is assumed when the link-based classifier returns a score ($score_l$) within a range of values having a lower bound threshold *t1* and an upper bound threshold *t2* (both thresholds are experimentally tuned). In the latter, the sheer number of words contained within a hypelink's local context designates whether there is room for the link-based classifier to produce a reliable estimation. Specifically, a number of surrounding words (i.e., *sw*) lower than a threshold k entails that the local context of a hyperlink is not representative enough.

In both cases, either the destination or the parent Web page classifier may be utilized so as to produce a more reliable estimation. When employed, the destination classifier acts a second step of assessment which produces a confidence score (i.e., $score_d$) for estimating the relevance of the page the hyperlink points to, based on its actual textual content (after downloading it). If $score_d$ exceeds threshold t, the page is considered relevant to the domain of interest. The downloaded Web page is stored locally, so as to avoid fetching it again in during the next fetching cycle. On the other hand, in order to avoid the destination classifier's burden hindering the focused crawler time performance, the employment of the parent Web page classifier provides an inexpensive solution capable of enhancing the link-based classifier estimation, which is really useful especially when a hyperlink's local context does not convey meaningful information. The parent classifier score (i.e., $score_p$) may be used either individually or in combination with $score_l$ produced by the link-based classifier. In the former case, if $score_p$ exceeds threshold t, the page is considered relevant to the domain of interest. In the latter, $score_l$ and $score_p$ are fused either in a static [4] or in a dynamic linear fashion, and their combination is used as the score for deciding the relevance of a hyperlink to the domain of interest. The static linear combination score (i.e., $score_{sc}$) is computed as

$$\text{score}_{sc} = a \times \text{score}_p + (1-a) \times \text{score}_l \quad (1)$$

where α is the relative weight assigned to the parent page score. When employed, the static combination produces a late fusion score which is always adjusted in a static way in accordance to $score_l$ and $score_p$.

On the other hand, the dynamic linear combination score (i.e. $score_{dc}$) is computed as

$$score_{dc} = \begin{cases} score_l \times ((1-\beta) - t + score_p), & \text{when} \quad \left(score_l \geq t \quad \text{AND} \quad score_p < t\right) \\ score_l \times ((1+\beta) - t + score_p), & \text{when} \quad \left(score_l \leq t \quad \text{AND} \quad score_p > t\right) \\ score_l & \text{otherwise} \end{cases} \quad (2)$$

with $0 \leq \beta < t$, where β is the weight determining the extent to which the dynamic late fusion process affects $score_l$ (higher β values entail less influence of $score_p$ to $score_l$). The dynamic linear combination approach is employed under the premise that there is discordance between the link-based and the parent page classifier estimation. Particularly, when $score_l$ is above threshold t and $score_p$ is below threshold t and vice versa, the dynamic combination technique adjusts negatively and positively, respectively, the score produced by the link-based classifier, aiming at producing an estimation conveying a more representative confidence score by taking advantage of the additional information included within the global context of the parent page. The core idea is to reduce or enhance $score_l$ depending on the distance of $score_p$ from threshold t; the greater the distance, the bigger the $score_l$ reduction or improvement, respectively. To the best of our knowledge this dynamic combination approach proposed is a new, not previously explored and tested strategy, which adjusts the confidence score based on a dynamic late fusion process, as opposed to the static process proposed in [4].

Based on the aforementioned rules governing the employment of the parent and the destination Web page classifiers, the link selection component supports 11 different methods for determining the hyperlink selection policy. These methods utilize combinations of the various classifiers discussed earlier, employed under different conditions based on the hyperlinks' local context strength and/or the destination network type. Table 1 summarizes the classifiers employed under each method investigated taking into account the constraints existing for their activation, whereas Table 2 illustrates the conditions under which each method under consideration operates. Method 1 employs only the link-based classifier regardless of the network type encountered, whereas for method 2 the link classifier is applied on hyperlinks pointing to the Surface Web and the destination Web page classifier is applied on Dark Web hyperlinks. Methods 3, 4, 5, 6, 7, and 8 all employ a two-step strategy, where the link-based classifier is accompanied by an additional classifier (from the pool of the available classifiers discussed earlier), activated when the crawl process encounters hyperlinks with insufficient local evidence (i.e., methods 3, 4, 5, and 6 are based on whether the link-based classifier produces a strong confidence score, while methods 7 and 8 depend on the number of words surrounding a hyperlink on a parent page). For every method implementing a two-step strategy, the additional classifier is enabled also depending on the destination network type. Methods 3, 5, and 6 employ the destination, the parent, and the static combination classifier,

Table 1 Summary of the 11 hyperlink selection methods proposed. The table lists the classifiers involved, the local context constraints (where applied), and the network type where each classifier is applied for each method under consideration

Methods	Classifiers					Local context constraints		Classifiers used per Network Type			
	L	P	D	SC_{PL}	DC_{PL}	$t_1 < \text{score}_l < t_2$	$\lvert sw \rvert \leq k$	Freenet	Tor	I2P	Surf.
1	x							L	L	L	L
2	x		x					D	D	D	L
3	x		x			x		L, D	L, D	L	L
4	x		x			x		L, D	L, D	L, D	L, D
5	x	x				x		L, P	L, P	L	L
6	x			x		x		L, SC_{PL}	L, SC_{PL}	L	L
7	x		x				x	L, D	L, D	L, D	L, D
8	x	x					x	L, P	L, P	L, P	L, P
9				x				SC_{PL}	SC_{PL}	SC_{PL}	SC_{PL}
10					x			DC_{PL}	DC_{PL}	DC_{PL}	DC_{PL}
11	x				x			DC_{PL}	DC_{PL}	DC_{PL}	L

L stands for the link-based classifier, *P* for the parent Web page classifier, *D* for the destination Web page classifier, SC_{PL} for the static linear combination classifier, DC_{PL} for the dynamic linear combination classifier (both for the static and the dynamic combination classifiers the link-based and the parent classifier are combined), *Surf* for the Surface Web. The comma separated entries for the classifiers used per network type, entail a two-step hyperlink selection strategy, where the second classifier is enabled based on the local context constraints

Table 2 Conditions under which each one of the 11 methods proposed operates

Methods	Conditions
1	$score_l \geq t$
2	$(score_d \geq t_1$ AND $link \in \{Tor, Freenet, I2P\})$ OR $(score_l \geq t_2$ AND $link \notin \{Tor, Freenet, I2P\})$
3	$score_l \geq t_2$ OR $(t_1 < score_l < t_2$ AND $link \in \{Tor, Freenet\}$ AND $score_d \geq t)$ OR $(t_1 < score_l < t_2$ AND $link \notin \{Tor, Freenet\}$ AND $score_l \geq t)$
4	$score_l \geq t_2$ OR $(t_1 < score_l < t_2$ AND $score_d \geq t)$
5	$score_l \geq t_2$ OR $(t_1 < score_l < t_2$ AND $link \in \{Tor, Freenet\}$ AND $score_p \geq t)$ OR $(t_1 < score_l < t_2$ AND $link \notin \{Tor, Freenet\}$ AND $score_l \geq t)$
6	$score_l \geq t_2$ OR $(t_1 < score_l < t_2$ AND $link \in \{Tor, Freenet\}$ AND $score_{sc} \geq t)$ OR $(t_1 < score_l < t_2$ AND $link \notin \{Tor, Freenet\}$ AND $score_l \geq t)$
7	$score = \begin{cases} score_l, & if \quad \lvert sw \rvert \geq 3 \\ score_d, & otherwise \end{cases}$
8	$score = \begin{cases} score_l, & if \quad \lvert sw \rvert \geq 3 \\ score_p, & otherwise \end{cases}$
9	$score_{sc} \geq t$
10	$score_{dc} \geq t$
11	$(score_l \geq t$ AND $link \notin \{Tor, Freenet, I2P\})$ $(score_{dc} \geq t$ AND $link \in \{Tor, Freenet, I2P\})$ OR

respectively, when Freenet and Tor hyperlinks are traversed, whereas methods 4 and 7 exploit the destination classifier and method 8 the parent classifier, regardless of a hyperlink's network type. Finally, methods 9 and 10 employ the static and the dynamic linear combination classifier, respectively, regardless of the destination network type, whereas method 11 applies the dynamic linear combination classifier only for Dark Web hyperlinks.

8 Evaluation

To assess the effectiveness of the proposed hybrid crawler focused on the HME domain, several experiments were performed based on Web pages from the Surface and the Dark Web, after investigating the 11 hyperlink selection methods proposed, which exploit the link-based classifier in conjunction with either the parent or the destination Web page classifier.

Given the large scale of crawling experiments, automatic relevance assessments are often employed based, for instance, on the lexical similarity of the textual content of the discovered Web pages to a detailed textual description of the topic, or on the confidence score of a classifier trained on the topic [15, 17]. Since this introduces some level of noise that may inadvertently affect the performance evaluation, and thus not allow us to fully gauge the effectiveness of the proposed approaches, the samples used for building the hybrid focused crawler classifiers were manually annotated by law enforcement agents and domain experts participating in HOMER using the following binary relevance scale:

- *relevant*: Web resources containing recipes which describe how to synthesize an HME or include other useful information, such as the properties of an explosive, or the demonstration of an HME
- *non-relevant*: Web resources not containing any information about HMEs or including weakly relevant information, such as, for instance, news articles about a bomb attack using HMEs, without listing any information regarding the HME

All three classifiers were built using supervised machine learning based on Support Vector Machines (SVMs) and were trained for the HME domain. For training the link-based classifier, 400 samples (105 positive, 295 negative) were used for building the classifier based on the local representation of the hyperlinks discussed in the "Frontier component" section. An SVM with an RBF kernel is employed, and class weight parameters are selected by performing fivefold cross-validation on the training set. On the other hand, both the parent and the destination Web page classifiers are trained based on 600 samples (250 positive, 350 negative) using the global textual representation of a Web page, as discussed in the "Frontier component" section. An SVM classifier is trained using an RBF kernel, while tenfold cross-validation is performed for selecting the class weight parameters

Next, this section describes the experimental set-up and then discusses the evaluation results.

9 Experiments

A set of four seed URLs was used in the experiments (1 Surface Web URL, 1 Tor URL, 1 I2P URL and 1 Freenet URL). The actual URLs are not provided so as to avoid the inclusion of potentially sensitive information, but are available upon request. These seed URLs were obtained in the aforementioned empirical study after submitting HME-related queries in several search engines of the Surface and the Dark Web, such as Yahoo!, Bing, DuckDuckGo, Ahmia, and Torch. The relatively small seed set is employed so as to keep the evaluation tractable investigating larger crawling depths, compared to our preliminary study. Starting from these seeds, a crawl at depth = 3 (i.e., the maximum distance allowed between seed and crawled pages) was performed in February 2017 and was locally stored.

The experiments conducted investigate the effectiveness of the 11 hyperlink selection methods proposed for the HME domain and have been categorized based on the relevance of the basic characteristics encountered in

each method. All experiments are performed for t values ranging from 0.5 to 0.9 at step 0.1. The first group of experiments includes methods employing basic link selection depending solely either on the link-based classifier or on the activation of the destination Web page classifier based on the network type encountered irrespective of other conditions (i.e., method 1 is based only on the link-based classifier, whereas method 2 enables the destination classifier when the hyperlink encountered points to the Dark Web). These two methods are used as a baseline for further comparisons with more sophisticated solutions, taking into account additional conditions governing the Surface and the Dark Web.

Furthermore, the second group of experiments explores the effectiveness of the methods employed to deal with the lack of sufficient local context around hyperlinks. To this end, it includes six methods (i.e., methods 3, 4, 5, 6, 7, and 8) which employ a two-step strategy based on whether there is strong evidence within a hyperlink's local context. For methods 3, 4, 5, and 6, in case the link-based classifier produces a confidence score within a range of values $t1$ and $t2$, $t1 < t2$ (for hyperlinks belonging to the network types accepted in each experiment), and irrespective of the value of threshold t, an additional Web page classifier is enabled. On the other hand, with regard to methods 7 and 8, the additional classifier is enabled only when the number of surrounding words around a hyperlink sw is less than a threshold k. For this set of experiments, thresholds t_1 and t_2 were set to 0.3 and 0.7, respectively, and k is set to 2 after experimental tuning. Finally, the third group of experiments investigates the effectiveness of the methods (i.e., methods 9, 10, and 11) employing either the static or the dynamic linear combination classifier. In particular, methods 9 and 10 employ the static and the dynamic combination classifier irrespective of the network type encountered, whereas method 11 applies the dynamic combination classifier for Dark Web hyperlinks. For this group of experiments, α is set to 0.25 and β is set to 0, after experimental tuning.

10 Results

This section provides the evaluation results of the experiments performed for assessing the effectiveness of the hybrid focused crawler for the HME domain based on their precision, recall, and F-measure, as well as the efficiency of the 11 hyperlink selection methods based on the number of Web page hits for the Surface and the Dark Web (i.e., each hit entails that a Web page has been visited and downloaded; thus, we consider that the time needed for completing a crawl increases linearly in relation to the number of the hits). Given that recall requires knowledge of all relevant pages on a given topic (an impossible task in the context of the Web), it is computed by manually designating a few representative pages on the HME domain and measuring what fraction of them are discovered by the crawler [3]. Here, we consider the set of unique relevant Web pages retrieved by the focused crawler when t is set to 0.5 (given that it corresponds to a superset for all other thresholds $t > 0.5$) after running crawls applying all 11 hyperlink selection approaches (i.e., 397 samples in total). The total number of unique Web pages used for the evaluation is 846 and corresponds to the total number of unique pages retrieved when t is set to 0.5 for all 11 experiments.

Tables 3, 4, and 5 present the results of the experiments evaluating the focused crawler effectiveness and efficiency for various values of threshold t at depth = 3. Specifically, Table 3 presents the results for the first group of experiments (i.e., baseline experiments), Table 3 illustrates the results for the second group of experiments (i.e., two-step strategies dealing with the lack of local evidence around a hyperlink), whereas Table 5 includes the results for the third set of experiments (i.e., utilizing the static or the dynamic linear combination methods). As expected, precision increases for higher values of threshold t in all 11 experiments, whereas recall demonstrates the opposite tradeoff. In general, the threshold value 0.7 appears to be a rather good compromise, since it achieves a high precision, while still maintaining a significant recall.

Investigating the first set of experiments results in concluding that both methods (i.e., methods 1 and 2) exhibit similar precision-related performance, with method 1 having slightly higher values than method 2 for lower threshold t values (i.e., $t < 0.7$) and vice versa. However, with regard to the recall performance, method 2 exhibits much higher values, which in turn,

Table 3 Focused crawler effectiveness and efficiency when the first group of hyperlink selection methods are employed

Methods		Threshold t				
		0.5	0.6	0.7	0.8	0.9
1	Precision	0.687	0.687	0.777	0804	0.913
	Recall	0.624	0.536	0.520	0.480	0.266
	F-measure	0.654	0.602	0.623	0.601	0.413
	Surface Web hits	203	192	176	173	49
	Dark Web hits	126	103	66	61	24
	Total hits	329	295	242	234	73
2	Precision	0.649	0.675	0.769	0.822	0.929
	Recall	0.896	0.764	0.711	0.622	0.365
	F-measure	0.753	0.717	0.739	0.708	0.525
	Surface Web hits	203	192	176	173	49
	Dark Web hits	4198	4198	4198	4198	4198
	Total hits	4401	4390	4374	4371	4147

Table 4 Focused crawler effectiveness and efficiency when the second group of hyperlink selection methods are employed

Methods		Threshold t				
		0.5	0.6	0.7	0.8	0.9
3	Precision	0.742	0.741	0.777	0.805	0.914
	Recall	0.619	0.538	0.523	0.482	0.269
	F-measure	0.675	0.624	0.625	0.603	0.416
	Surface Web hits	203	192	176	173	49
	Dark Web hits	132	122	111	106	69
	Total hits	335	314	287	279	118
4	Precision	0.768	0.770	0.766	0.779	0.874
	Recall	0.596	0.579	0.563	0.510	0.282
	F-measure	0.671	0.661	0.649	0.617	0.426
	Surface Web hits	498	496	496	464	227
	Dark Web hits	138	136	136	131	94
	Total hits	636	632	632	595	321
5	Precision	0.743	0.740	0.777	0.804	0.913
	Recall	0.617	0.536	0.520	0.480	0.266
	F-measure	0.674	0.622	0.623	0.601	0.413
	Surface Web hits	203	192	176	173	49
	Dark Web hits	86	77	66	61	24
	Total hits	289	269	242	234	73
6	Precision	0.698	0.740	0.777	0.804	0.913
	Recall	0.617	0.837	0.520	0.480	0.266
	F-measure	0.655	0.786	0.623	0.601	0.413
	Surface Web hits	203	192	176	173	49
	Dark Web hits	111	77	66	61	24
	Total hits	314	269	242	234	73
7	Precision	0.687	0.687	0.777	0.804	0.913
	Recall	0.624	0.536	0.520	0.480	0.266
	F-measure	0.654	0.602	0.623	0.601	0.413
	Surface Web hits	513	486	470	451	231
	Dark Web hits	127	67	62	62	25
	Total hits	640	537	513	513	256
8	Precision	0.674	0.672	0.765	0.791	0.913
	Recall	0.624	0.536	0.520	0.480	0.266
	F-measure	0.648	0.596	0.619	0.597	0.413
	Surface Web hits	211	200	176	173	49
	Dark Web hits	126	103	66	61	24
	Total hits	337	303	242	234	73

Table 5 Focused crawler effectiveness and efficiency when the third group of hyperlink selection methods are employed

Methods		Threshold t				
		0.5	0.6	0.7	0.8	0.9
9	Precision	0.700	0.770	0.798	0.780	–
	Recall	0.599	0.561	0.513	0.251	0.000
	F-measure	0.646	0.649	0.624	0.380	–
	Surface Web hits	198	186	172	161	0
	Dark Web hits	102	62	44	0	0
	Total hits	300	248	216	161	0
10	Precision	0.768	0.759	0.728	0.929	–
	Recall	0.589	0.439	0.292	0.066	0.000
	F-measure	0.667	0.556	0.417	0.123	–
	Surface Web hits	190	176	163	48	0
	Dark Web hits	62	0	0	0	0
	Total hits	252	176	163	48	0
11	Precision	0.759	0.745	0.771	0.807	0.903
	Recall	0.599	0.416	0.401	0.371	0.213
	F-measure	0.670	0.534	0.528	0.508	0.345
	Surface Web hits	203	192	176	173	49
	Dark Web hits	62	0	0	0	0
	Total hits	265	192	176	173	49

as expected, affects the F-measure score as well. This behavior is explained by the fact that method 2 applies the destination Web page classifier for every hyperlink pointing to a Dark Web page, and as a result, the breadth of the relevant pages discovered is increased significantly. However, this significant improvement on the recall and F-measure score comes with a big overhead of downloading every single Dark Web page encountered through the crawling process, among which are several pages non-relevant to the HME domain (e.g., method 2 visits 4198 Dark Web pages, whereas method 1 only 126 when t is set to 0.5). Therefore, the additional methods proposed potentially incurring less overhead compared to method 2 are explored in order to reach a good compromise, which improves the effectiveness of method 1 without adding a significant overhead in terms of the time needed for running a crawl.

Examining the second group of experiments reveals the slightly better performance in terms of precision and recall for methods 3, 4, and 5 (they all exhibit similar performance) compared to the remaining methods (i.e., methods 6, 7, and 8). The three most effective methods improve precision and F-measure for low threshold t values (i.e., $t < 0.7$) without decreasing recall compared to the baseline method 1 (i.e., using only the link-based classifier) whereas the remaining methods exhibit similar results with method 1. Among the three most effective methods, the one exhibiting the highest efficiency is method 5, as a result of employing a parent Web page classifier as the additional classifier activated, which entails that there is no extra overhead for downloading the destination Web page as opposed to methods 3 and 4.

At the same time, when compared with method 2, precision improves for threshold t values with $t < 0.8$ and the F-measure score is still high, considering the much lower overhead method 5 incurs to the whole crawling process (e.g., method 2 visits 4112 additional pages when t is set to 0.5 compared to method 5).

Furthermore, regarding the third set of experiments, method 9 exhibits similar precision results with the baseline methods 1 and 2 for threshold t values with $t < 0.9$, whereas for methods 10 and 11, precision improves significantly for almost all threshold values. Nevertheless, recall drops significantly, especially for higher threshold t values (i.e., $t > 0.5$), which entails that the novel dynamic linear combination strategy proposed has the potential of discovering relevant pages with higher accuracy; however, it fails to expand the knowledge to the HME domain. Additionally, methods 10 and 11 significantly outperform method 2 in terms of efficiency, since they do not require fetching the content present in destination Web pages (e.g., method 2 downloads 3904 and 4140 additional Web when t is set to 0.5 in comparison to methods 10 and 11, respectively).

To sum up, the experiments conducted reveal the potential of the proposed methods to improve the effectiveness of the focused crawler, without decreasing its efficiency. Selecting a method for performing a crawl depends on the desirable goal. For example, for precision-oriented crawls in lower threshold t values, methods 10 and 11 provide reliable results, whereas for recall-oriented crawls, method 2 outperforms by far the rest of the methods; however, this high recall value comes with a significant cost in terms of time performance. For crawls aiming at accomplishing a balanced tradeoff among precision and recall, method 5 appears to be the most suitable solution.

11 Conclusions

This work proposed a hybrid focused crawler capable of seamlessly following hyperlinks pointing to resources hosted on the Surface Web and several darknets of the Dark Web (i.e., Tor, I2P, and Freenet) during a single crawl. It employs a classifier-guided approach for selecting the hyperlinks to follow which combines three classifiers: (i) a link-based classifier exploiting the local context of hyperlinks in their parent page, (ii) a parent page classifier taking advantage of the actual textual content of the parent page of a hyperlink, and (iii) a destination page classifier exploiting the global context of the page a hyperlink points to. These three classifiers are employed by the 11 hyperlink selection methods supported by the link selection component of the focused crawler, including the novel dynamic linear combination approach proposed. Each method employs a different combination of the available classifiers, where the link-based classifier is accompanied by the parent or the destination Web page classifier, depending on conditions related to the destination network and/or the existence of strong evidence within the local context of a hyperlink.

The proposed hybrid focused crawler can be used by LEAs and Intelligence Agencies as a significant tool in the fight against terrorism and extreme violence. Although this work focused on the discovery of HME related content, the same framework can be reused for discovering information related to any topic of interest, including other illegal and terrorist-related content on the Web such as drugs, child pornography, money counterfeiting, human trafficking, or terrorist radicalization, simply by building appropriate classifiers in each case. Future work will investigate the incorporation of mechanisms based on machine learning techniques that will be used so as not only to avoid the existing detection techniques from the sites of interest but also to readjust the crawler's behavior to avoid any additional countermeasures that may be applied. Of course, the legal aspects of the above techniques should always be considered and the respective legislation should be followed.

12 Endnotes

[1]https://www.torproject.org/
[2]https://geti2p.net/en/
[3]https://freenetproject.org/
[4]https://ahmia.fi/ and
http://msydqstlz2kzerdg.onion/
[5]http://xmh57jrzrnw6insl.onion/
[6]http://homer-project.eu/
[7]https://nutch.apache.org/

Acknowledgements
This work was supported by the TENSOR (700024) and HOMER (312388) projects partially funded by the European Commission.

Competing interests
The authors declare that they have no competing interests.

13 Publisher's Note

Received: 6 March 2017 Accepted: 21 June 2017
Published online: 04 July 2017

References

1. V Ciancaglini, M Balduzzi, R McArdle, M Rösler, *The Deep Web, Trend Micro*, 2015
2. C Iliou, G Kalpakis, T Tsikrika, S Vrochidis, I Kompatsiaris, in 11th International Conference on Availability, Reliability and Security (ARES). Hybrid Focused Crawling for Homemade Explosives Discovery on Surface and Dark Web (IEEE, 2016), p. 229–234.
3. C Olston, M Najork. Web crawling. Foundations and Trend in Information Retrieval **4**(3), 175–246 (2010).

4. G Pant, P Srinivasan, Link contexts in classifier-guided topical crawlers. IEEE Transactions on Knowledge and Data Engineering **18**, 107–122 (2006). doi:10.1109/TKDE.2006.12
5. Project Artemis – OSINT Activities on Deep Web, infosecinstitute.com, July 2013. http://resources.infosecinstitute.com/project-artemis-osint-activities-on-deep-web/. Accessed 20 Dec 2016.
6. Memex Project (Domain-Specific Search) Open Catalog. https://opencatalog.darpa.mil/MEMEX.html. Accessed 20 Dec 2016.
7. T Fu, A Abbasi, H Chen (2010). A focused crawler for Dark Web forums. *Journal of the American Society for Information Science and Technology* **61**, 1213–1231. doi: 10.1002/asi.21323.
8. A Biryukov, I Pustogarov, F Thill, RP Weinmann, in 1st International Workshop on Big Data Analytics for Security (DASec) of the 34th International Conference on Distributed Computing Systems Workshops (ICDCS Workshops). Content and Popularity Analysis of Tor Hidden Services (IEEE, 2014), pp. 188–193. doi: 10.1109/ICDCSW.2014.20.
9. H Chen (2011). Dark Web: Exploring and Data Mining the Dark Side of the Web, vol. 30, Springer Science & Business Media..
10. H Chen, W Chung, J Qin, E Reid, M Sageman, G Weimann (2008). Uncovering the Dark Web: a case study of Jihad on the Web. Journal of the American Society for Information Science and Technology **59**, 1347–1359.
11. G Kalpakis, T Tsikrika, F Markatopoulou, N Pittaras, S Vrochidis, V Mezaris, I Patras, I Kompatsiaris, in 10th International International Conference on Availability, Reliability and Security (ARES). Concept Detection on Multimedia Web Resources about Home Made Explosives (IEEE, 2015), p. 632–641. doi: 10.1109/ARES.2015.85.
12. T Tsikrika, G Kalpakis, S Vrochidis, I Kompatsiaris, I Paraskakis, I Kavasidis, J Middleton, U Williamson, in 10th International Conference on Availability, Reliability and Security (ARES). A framework for the discovery, analysis, and retrieval of multimedia homemade explosives information on the Web (IEEE, 2015), p. 601-610. doi:10.1109/ARES.2015.86.
13. G Kalpakis, T Tsikrika, C Iliou, T Mironidis, S Vrochidis, J Middleton, U Williamson, I Kompatsiaris, in 18th International Conference on Human-Computer Interaction (HCI). Interactive Discovery and Retrieval of Web Resources Containing Home Made Explosive Recipes (press, 2016)
14. T Tsikrika, A Moumtzidou, S Vrochidis, I Kompatsiaris, Focussed crawling of environmental Web resources based on the combination of multimedia evidence. Multimedia Tools and Applications **75**(3), 1563–1587 (2016). doi:10.1007/s11042-015-2624-3
15. G Pant, P Srinivasan, Learning to crawl: comparing classification schemes. ACM Transactions on Information Systems (TOIS) **23**(4), 430–462 (2005). doi:10.1145/1095872.1095875
16. M Hersovici, M Jacovi, YS Maarek, D Pelleg, M Shtalhaim, S Ur, The shark-search algorithm—an application: tailored Web site mapping. Journal Computer Networks and ISDN Systems **30**(1), 317–326 (1998)
17. P Srinivasan, F Menczer, G Pant, A general evaluation framework for topical crawlers. Journal Information Retrieve **8**(3), 417–447 (2005). doi:10.1007/s10791-005-6993-5