Interactive Discovery and Retrieval of Web Resources Containing Home Made Explosive Recipes

George Kalpakis[1(✉)], Theodora Tsikrika[1], Christos Iliou[1], Thodoris Mironidis[1],
Stefanos Vrochidis[1], Jonathan Middleton[2], Una Williamson[2],
and Ioannis Kompatsiaris[1]

[1] Information Technology Institute, CERTH, Thessaloniki, Greece
{kalpakis,theodora.tsikrika,iliouchristos,mironidis,
stefanos,ikom}@iti.gr
[2] Police Service Northern Ireland, Belfast, UK
{jonathan.middleton,una.williamson}@psni.pnn.police.uk

Abstract. This work investigates the effectiveness of a novel interactive search engine in the context of discovering and retrieving Web resources containing recipes for synthesizing Home Made Explosives (HMEs). The discovery of HME Web resources both on Surface and Dark Web is addressed as a domain-specific search problem; the architecture of the search engine is based on a hybrid infrastructure that combines two different approaches: (i) a Web crawler focused on the HME domain; (ii) the submission of HME domain-specific queries to general-purpose search engines. Both approaches are accompanied by a user-initiated post-processing classification for reducing the potential noise in the discovery results. The design of the application is built based on the distinctive nature of law enforcement agency user requirements, which dictate the interactive discovery and the accurate filtering of Web resources containing HME recipes. The experiments evaluating the effectiveness of our application demonstrate its satisfactory performance, which in turn indicates the significant potential of the adopted approaches on the HME domain.

Keywords: Interactive search engine · Homemade explosives · Dark web

1 Introduction

The large number of terrorist attacks that have taken place worldwide during the past 20 years has put increasing pressure to law enforcement agencies (LEAs), so as to uphold the rule of law by raising their awareness against terrorist groups to the highest level possible. The recent escalation of terrorist attacks clearly indicates that the war against terrorism should be fought with all the available means, including alternative solutions offered through the continuous advancement of technology. At the same time, the rapid growth of broadband technologies, along with the abundance of online resources for storing content, has resulted in the proliferation of the information being shared globally. This growth has facilitated the diffusion of knowledge in many different domains, nevertheless it has resulted in sharing information online that poses a threat to the society,

T. Tryfonas (Ed.): HAS 2016, LNCS 9750, pp. 221–233, 2016.
DOI: 10.1007/978-3-319-39381-0_20

such as material that can be used for supporting acts of terrorism, including information for the manufacture and use of homemade explosives (HMEs) which can be exploited for subversive use. The availability of such material on the Web provides the subversive with the ability to thoroughly study the process of synthesizing HMEs, using common household goods and easy to purchase items. Hence, it is vital for LEAs to exploit technologies that will enable them to cope with this threat by automatically identifying resources with HME information.

To meet this challenge, this work proposes a novel interactive search engine for the discovery and retrieval of Web resources including HME content, with particular focus on HME recipe information. Such information, present both on Surface and Dark Web, can be found on various types of Web resources, such as Web pages, forums, and social media posts. The proposed application enables the discovery of HME content through a user friendly interface. The interactive search engine is built based on the continuous need for LEAs to discover HME information on the Web and to filter the most significant resources by interacting with tools capable of accurately distinguishing the most relevant resources within the HME domain. It employs a hybrid model consisting of four major components that provide several advanced facilities: (i) a **Web Crawling** component focused on the HME domain (ii) a **Querying** component which submits HME domain-specific queries to general-purpose search engines, (iii) a **Post-Processing Classification** component which reduces the potential noise of the discovery results, and (iv) an **Interactive Graphical User Interface** which facilitates the user communication with the search engine's main components.

The main contribution of this work is the integration of domain-specific technologies in a novel interactive search engine for the discovery and retrieval of heterogeneous Web resources containing HME information, as well as the adaptation of domain-specific search technologies in the context of the HME domain. This interactive search engine has been developed in a user-driven manner in collaboration with and based on the requirements of LEA personnel, and provides access to HME crawling and querying tools via a unified Graphical User Interface tailored to facilitating the intelligence gathering process. Additionally, this application provides a combination of domain-specific tools applied both to Surface and Dark Web, where illegal and potentially harmful information is usually stored. To the best our knowledge, this is the first attempt to develop an interactive search engine designed to facilitate the discovery of HME information that combines search and crawling capabilities.

2 Use Cases and Requirements

This section discusses the end-user requirements of the proposed interactive search engine that are elicited based on appropriate use cases for HME discovery. The use cases have been provided by law enforcement and security agents in the context of HOMER project[1] who have offered guidance for a user-oriented development.

[1] http://www.homer-project.eu/.

Use Case 1. Search for HME-related Content in an Extremist Forum: LEAs need to detect user posts present in specific Web forums in Surface and Dark Web discussing the process of synthesizing an HME recipe, and providing feedback on how to use such a product in terrorist activities. For example, consider the case that a group of users, members of a forum identified for its extremist character, discuss their intention of manufacturing an HME for using it in an imminent terrorist attack.

Use Case 2. Search the Web for Information Related to an Explosive: LEAs need to constantly discover new sources with HME content and monitor the advancements in the manufacture of HMEs by submitting keyword-based queries enhanced with a mechanism for their automatic formulation or expansion. They, also, need to get results ranked by their relevance to the HME domain. Such information is usually being shared online and indexed by general-purpose search engines of Surface and/or Dark Web. For instance, consider the case there is intelligence that a specific substance not previously used is currently being considered for subversive use.

Both use cases reflect the challenge of developing an interactive search engine for supporting searches for HME content on Surface and Dark Web. Table 1 presents the core requirements of the interactive search engine as emerged by the two use cases.

Table 1. Interactive search engine requirements

Requirements	Description	Use Case
R1 – Traverse the Web looking for HME content	Search for Web resources via a classifier-guided crawler focused on the HME domain	UC1
R2 – Submit keyword-based automatically formulated queries for finding HME-related content	Search for HME Web resources by submitting automatically formulated queries to general-purpose search engines based on keywords applied to a set of domain-specific query patterns	UC2
R3 – Submit keyword-based automatically expanded queries	Search for HME Web resources by expanding keyword-based queries using a set of query expansion rules and submitting them to general-purpose search engines	UC2
R4 – Filter the returned results	Re-rank the return results by estimating their relevance to the HME-domain based on classification	UC2
R5 – Interactive User Interface	Provide an interactive Graphical User Interface for running, configuring and parameterizing the different modes of search provided	UC1, UC2
R6 – Perform searches for HMEs in Dark Web	Develop a tool supporting focused crawling in Dark Web anonymous networks, and query submitting in general-purpose search engines for Dark Web resources.	UC1, UC2

3 Related Work

This section first discusses the state-of-the-art approaches in the field of domain-specific search and then examines the major research efforts for discovering terrorist or extremist-related content on Web resources.

Domain-specific search can be considered as a well-established research area, since a considerable number of techniques have been developed to tackle this issue. The main methodologies used for the implementation of domain-specific search tools for the discovery of Web resources on a given topic can be divided into two categories: (i) methods based on focused Web crawling and (ii) methods based on the indexes and search infrastructures of existing general-purpose search engines. For the first category of methods, the state-of-the-art approaches [1] adopt classifier-guided crawling strategies based on supervised machine learning methods that rely on (i) the hyperlinks' local context, and/or (ii) global evidence associated with the entire parent page. For the second category of methods, existing general-purpose search engines (e.g., Google, Yahoo!, etc.) are employed and domain-specific queries generated semi-automatically or fully automatically are submitted to them. The queries generated semi-automatically are based on abstract query patterns that can then be instantiated into multiple query instances corresponding to sequences of domain-specific keywords and attributes [2]. More complex domain-oriented queries can be automatically generated by applying machine learning techniques in order to extract terms referred to as "keyword spices" [3, 4]. However none of these approaches have been applied and adapted in the context of the HME domain.

In the field of discovering terrorist-related information on the Web, a thorough research [5] has paved the way for the development tools that aim to collect and analyze Web content generated by international terrorist groups. To this end, the most comprehensive suite of text mining and Web mining tools for performing link and content, analysis has been developed in the context of the Dark Web project at the University of Arizona [6]. However, this project has addressed the whole breadth of terrorist and extremist content, rather than HME information, as done here. Furthermore, research efforts have been conducted for performing focused crawling in Dark Web forums moderated by extremist groups by using a human-assisted accessibility approach [7]. Moreover, another methodology, applied in a set of Jihad Web sites, has been proposed for collecting and analyzing Dark Web information for aiding the process of intelligence gathering [8]. Again, they deal with a wider scope of information, rather than only with HMEs. Additionally, a work related to HMEs presents a mechanism for automatically identifying the relevance of already discovered multimedia content (videos/images) to the HME domain based on concept detection [9]; however, this work deals with the identification of HME-related content in multimedia, rather than with the discovery and retrieval of such information from the Web. Furthermore, a relevant research effort presents a framework corresponding to a Knowledge Management Platform for managing the discovery, analysis and retrieval of HME-related content [10]; nevertheless, this work is mainly concerned with the general framework's architecture focusing on the HME knowledge management, rather than on the discovery of HME information from the Web. Contrary to the aforementioned papers, this work proposes a novel interactive search engine that focuses on HME information both on Surface and Dark Web, based on an innovative hybrid infrastructure, which consists not only of an interactive search mechanism, but also of a complementary online crawling tool, capable of traversing the Web on user demand. The search engine exploits domain-specific approaches and is being developed in a user-driven manner with the goal to fulfill the requirements of its end users.

4 HME Search Engine Architecture

This section provides a high level overview of the interactive search engine architecture for the discovery of Web resources with HMEs, which is designed based on the user requirements described earlier. As shown in Fig. 1, the underlying fundamental concept behind the adopted methodology in our application is the use of an interactive interface (R5) that provides access to different functionalities. The search engine enables the users (i.e. police and anti-terrorism/intelligence officers) to perform customized searches both on Surface and Dark Web (R6) for HME-related content by taking advantage of a hybrid model based both on focused crawlers (R1) and general-purpose search engines (R2, R3). The results of the discovery procedure go through a classification process on user demand for filtering out the potential noise (R4).

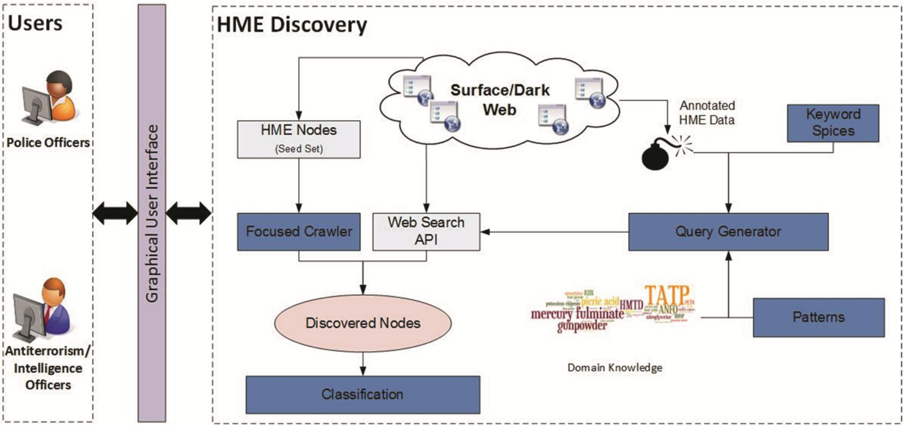


Fig. 1. Interactive search engine architecture

Our application's discovery and retrieval mechanism relies on a hybrid architecture, which combines into a joint workflow (i) a crawling facility where a Web crawler focused on the HME domain, starting from a predefined set of seed pages, performs selective traversal of Web resources by estimating their relevance to the HME domain based on supervised machine learning classifiers (R1), and (ii) a querying facility where HME domain-specific queries are submitted to general-purpose search engines; such queries are either formulated using query patterns in conjunction with domain knowledge (R2), or are expanded based on the methodology of “keyword spices” (i.e. domain-specific keywords generated with the aid of supervised machine learning techniques) (R3) [8]. These two methods are complementary, since the latter aims to exploit the large coverage of existing indexes containing Web resources already crawled in the very large scale by general-purpose search engines, as well as the search infrastructures they provide, while the former aims to address the inherent difficulties in the generation of effective domain-specific queries that would lead to the discovery of relevant Web resources, and also to go beyond what is already covered by existing indexes, by identifying new Web resources relevant to the HME domain that have not yet been indexed by a general-purpose search engine. Also, this hybrid infrastructure goes beyond exploring and discovering resources found typically on Surface Web. Both the crawling and the querying facility are capable of discovering HME resources on Dark Web (R6), whose nature facilitates the proliferation of information used for illegal and terrorist activities. Moreover, the use of focused crawling avoids dependencies on external services (i.e. existing search engines), thus ensuring the long-term viability of the application. Finally, post-processing filtering is performed on user demand based on a classification process (R4) for reducing the potential noise in the results obtained by both discovery approaches.

5 Search Engine Components

This section provides a detailed overview of the major components of the interactive search engine for the discovery and retrieval of HME information.

5.1 Focused Crawler

Our application employs a classifier-guided focused crawling approach for the discovery of HME Web resources, by starting from a set of relevant seed pages provided by the user. To this end, it estimates the relevance of a hyperlink to an unvisited resource based on its local context. Motivated by the results of an empirical study performed with the support of HME experts in the context of HOMER project indicating that the anchor text of hyperlinks leading to HME information often contains HME-related terms (e.g. the name of the HME), and also that the URL could also be informative to some extent, since it may contain relevant information (e.g. the name of the HME), we follow recent research [11] and represent the local context of each hyperlink using: (i) its anchor text, (ii) a text window of x characters (e.g. x = 50) surrounding the anchor text that does not overlap with the anchor text of adjacent links, and (iii) the terms extracted from the URL. Each sample 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 a supervised machine learning approach based on Support Vector Machines (SVMs), given their demonstrated effectiveness in such applications [12]. The confidence score for each hyperlink is obtained by applying a trained classifier on its feature vector, and the page pointed by the hyperlink is fetched if its score is above a given threshold.

The developed focused crawler is based on a customized version of Apache Nutch[2] (version 1.9). It has been configured so that it can be set to traverse several darknets present in Dark Web, and specifically the most popular anonymous networks, namely Tor[3], I2P[4] and Freenet[5]. Necessary for supporting crawling in Dark Web, is to enable the Tor, I2P and Freenet services on the machine running the crawler.

5.2 General-Purpose Web Search Engine Querying Component

This component submits domain-specific queries to existing general-purpose search engines both on Surface and Dark Web. Currently, for Surface Web, the Yahoo! BOSS API[6] is employed, but our tool can be easily extended so as to support additional APIs (provided by Google, Bing, etc.), whereas concerning the Dark Web, Duck Duck Go[7] and Ahmia[8] search engines are supported by executing queries on their Web interface (since no API is available) and parsing the returned results. The domain-specific queries are generated in a semi-automatic, or fully automatic fashion.

The semi-automatic approach requires the availability of an initial set of seed queries that can be processed (manually or/and automatically) so as to mine abstract query patterns that can then instantiated into multiple (concrete) query instances corresponding to sequences of domain-specific keywords [2, 10]. Here, an initial set of 45 queries that was formulated by law enforcement agents and was used for the successful discovery of HME Web resources through general-purpose search engines, is used as the seed set of queries. Once the keywords appearing in all queries are mapped to discrete concepts, then in every query in the initial seed set, the keywords are replaced by the respective concepts they are mapped to; for example, the query “preparation anfo” becomes “action explosive”. This results in producing a set of discrete patterns that can be used for automatic query generation (18 explosive-related patterns were produced). For example, the pattern “action explosive”, where “action” corresponds to keywords such as “how to make”, “preparation”, etc., may be instantiated to several different queries for each of the explosives of interest. Hence, once a user expresses interest in discovering HME Web resources about a given explosive, they can select query patterns containing the concept “explosive”, and these patterns will be automatically instantiated for that particular explosive, and will be submitted in parallel to the search engine; the results of all these queries will be merged before being presented.

The automatic approach aims to generate high precision and high recall queries in the HME domain. Based on machine learning techniques, it generates specific (Boolean)

[2] http://nutch.apache.org/.
[3] https://www.torproject.org/.
[4] https://geti2p.net/.
[5] https://freenetproject.org/.
[6] https://developer.yahoo.com/boss/search/.
[7] https://duckduckgo.com/.
[8] https://ahmia.fi/search/.

expressions (referred to as “keyword spices”) [4] that aim to characterize in an effective manner the HME domain. These expressions are then used for expanding (simple) domain-related queries; these expanded queries are subsequently submitted to a general-purpose search engine with the goal of improving the effectiveness of the initial (unexpanded) queries. The methodology implemented involves splitting an annotated set of Web resources into two subsets (for training and validation respectively), constructing a decision tree based on the training set and applying a decision tree learning algorithm for discovering the keyword spices [13]. Then these initial keyword spices are iteratively simplified by removing keywords in case their removal increases the F-measure of the validation set. The completion of this process results in generating the final set of Boolean expressions (8 expansion expressions were generated), which can be used for expanding simple queries [10]. An indicative example of such an expression generated is the following: *powder OR ingredient OR explosive.*

5.3 Post-processing Classification Component

The resources discovered through focused crawling and search engine querying are then classified based on their textual content. A text-based classifier is trained on a set of Web resources annotated as relevant or non-relevant 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 the frequency of term t in sample d, normalized by the maximum frequency of any t in that sample, and idf(t) is the inverse document frequency of term t in the collection of samples. Then, an SVM [14] classifier is trained using an RBF kernel, while 10-fold cross-validation is performed for selecting the class weight parameters. At query time the classifier, if initiated by the user, is capable of classifying the search engine and/or the focused crawling results as relevant or non-relevant to the HME domain. The classifier is implemented using the libraries of the Weka[9] machine learning software.

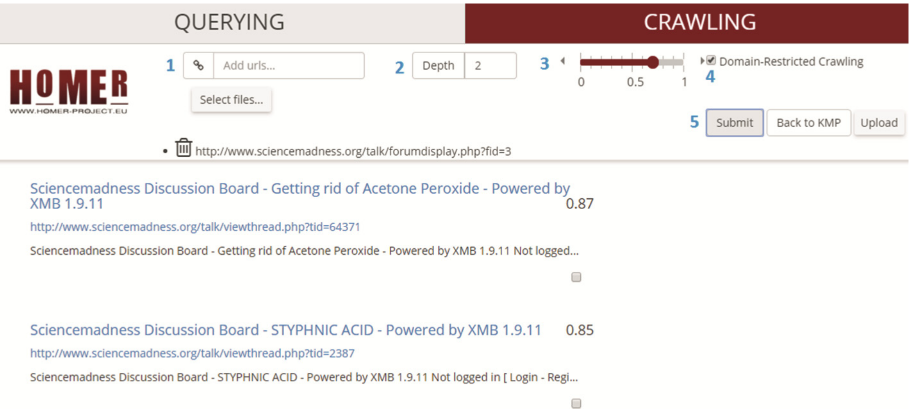
5.4 Graphical User Interface

The three aforementioned components of our application, namely the querying, the focused crawling and the classification components, are accessible via a Web-based intuitive Graphical User Interface (GUI) with a minimalist design aiming to provide a usable, flexible and consistent environment for the user, emphasizing in their interactive communication with the application. The GUI employs a responsive tabbed design allowing both the focused crawling and the querying facility to fully adapt to the screen size, using tabs as a navigational widget for switching between them.

Aiming to facilitate user engagement, the GUI provides a parametrized environment for running each one of these facilities. Specifically, for a focused crawling task (see Fig. 2), the following parameters can be configured: (i) the set of seed URLs constituting the starting points of the crawl, (ii) the score threshold that determines the necessary relevance

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

value that a URL needs to exceed in order to be accepted by the crawler classifier, (iii) the crawl depth (i.e. the maximum distance allowed between the seed pages and the crawled pages), and (iv) the option to perform a domain-restricted crawl (i.e. the crawler is allowed to follow hyperlinks belonging only to the same domain name(s) of the URL(s) present in the seed URL list). Additionally, for running a keyword-based query (see Fig. 3), three interaction modes are provided: (i) a free text mode for manually submitting queries, (ii) a semi-automatic mode for submitting queries taking advantage of 18 explosive-related query patterns, and (iii) an automatic mode for expanding the queries with one of the 8 available keyword spices. In both cases, a crawl or a querying run may be followed by a user-initiated post-processing classification process which re-ranks the respective returned results.



Depth 2

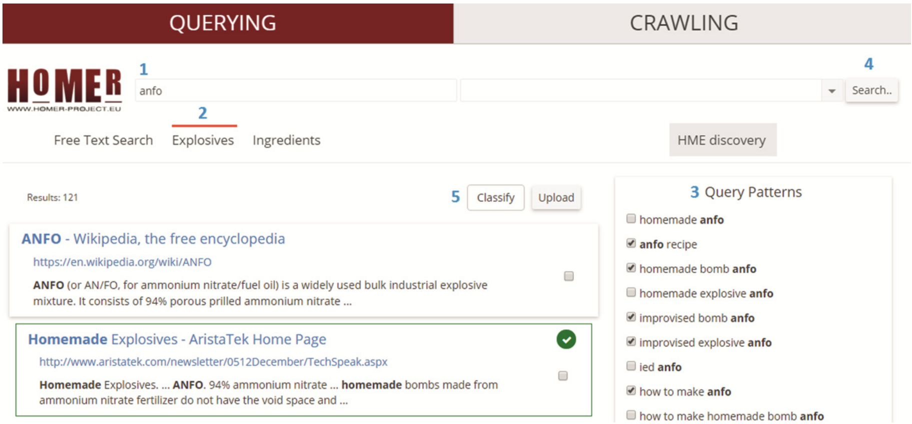
☑ Domain-Restricted Crawling

- http://www.sciencemadness.org/talk/forumdisplay.php?fid=3

Sciencemadness Discussion Board - Getting rid of Acetone Peroxide - Powered by XMB 1.9.11 0.87
http://www.sciencemadness.org/talk/viewthread.php?tid=64371
Sciencemadness Discussion Board - Getting rid of Acetone Peroxide - Powered by XMB 1.9.11 Not logged...
□

Sciencemadness Discussion Board - STYPHNIC ACID - Powered by XMB 1.9.11 0.85
http://www.sciencemadness.org/talk/viewthread.php?tid=2387
Sciencemadness Discussion Board - STYPHNIC ACID - Powered by XMB 1.9.11 Not logged in [Login - Regi...
□

Fig. 2. Focused crawling facility interface



anfo

Results: 121

ANFO - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/ANFO
ANFO (or AN/FO, for ammonium nitrate/fuel oil) is a widely used bulk industrial explosive mixture. It consists of 94% porous prilled ammonium nitrate ...
□

Homemade Explosives - AristaTek Home Page ✓
http://www.aristatek.com/newsletter/0512December/TechSpeak.aspx
Homemade Explosives. ... **ANFO**. 94% ammonium nitrate ... **homemade** bombs made from ammonium nitrate fertilizer do not have the void space and ...
□

Query Patterns
- □ homemade **anfo**
- ☑ **anfo** recipe
- ☑ homemade bomb **anfo**
- □ homemade explosive **anfo**
- ☑ improvised bomb **anfo**
- ☑ improvised explosive **anfo**
- □ ied **anfo**
- ☑ how to make **anfo**
- □ how to make homemade bomb **anfo**

Fig. 3. Querying facility interface

6 User Interaction Modes

This section presents the modes of interaction when using the proposed search engine for performing the tasks described in the use cases discussed in Sect. 2.

Interaction Mode 1. Search a Forum Looking for HME Content. In this case, the goal is to perform a domain-restricted crawl on forums where people are discussing about HME recipes, so as to discover posts related to synthesizing HMEs. Figure 2 illustrates the crawling facility interface. For performing a crawl, the user provides the seed URL(s) (in this case, a forum[10] Web page representing a section including topics related to HMEs) (1), selects the crawl depth (it is set to 2) (2), sets the crawler classifier's threshold (it is set to 0.7) (3), selects whether they wish to perform a domain-restricted crawl or not (a domain-restricted crawl is performed) (4), and initiates the crawl process (5). For the whole duration of the crawl, the interface provides continuous feedback updating the user about the process's progress. After the crawl is completed, the results revealing several HME forum posts are presented in descending order of their relevance to the HME domain.

Interaction Mode 2. Search the Web looking for an Explosive. In this scenario, the intention is to discover Web resources containing information about HMEs by submitting pattern-based queries to general-purpose search engines. Figure 3 illustrates the querying facility interface. For performing a pattern-based querying on explosives, the user enters the desirable keyword(s) representing the explosive of interest (in this case, the keyword provided is anfo) (1), indicates that they want to use the explosive-related patterns (2), selects one or more of the available patterns (3), and initiates the querying (4). When the querying is completed, the results, after being merged, are presented to the user. A result is depicted in a colored box depending on whether the Web page it represents has been previously annotated by domain experts.

There are 2 color variations: (a) green depicts relevant results and (b) red depicts non-relevant results. After performing the querying, the user has the option to initiate the classification process (5). When the classification is completed, the results are re-ranked and presented along with their score of relevance to the HME domain.

7 Evaluation

This section provides the evaluation results of the experiments performed for assessing the effectiveness of the interactive search engine's major facilities, namely the focused crawling and the querying for discovering HME-related content.

First, we present the evaluation results for crawling emphasizing in measuring precision (recall requires knowledge of all relevant pages on a given topic, an impossible task in the context of the Web) for different thresholds. In this case, a set of 254 pages fetched by the focused crawler, when the threshold t is set to 0.5 has been assessed based on a

[10] http://www.sciencemadness.org/.

two-point relevance scale, characterizing the retrieved resources as being relevant (i.e. resources describing HME recipes, explosive properties etc.) or non-relevant to the HME domain. The results of the experiments for various values of threshold *t* at depth = 3 are presented in Table 2. As expected, precision increases for higher values of threshold *t*. In particular, the difference between threshold values 0.6 and 0.7 is quite significant, with precision improving significantly.

Table 2. Focused crawling evaluation results

	Threshold t				
	0.5	0.6	0.7	0.8	0.9
Precision	0.52	0.54	0.73	0.73	0.74

With respect to the query formulation module, we provide a comparison for the performance of the different techniques involved, including domain-specific keywords and enhanced queries combining keywords with patterns or keyword spices. We emphasize on mean average precision, while the recall performance is also discussed. For evaluating the querying facility, we utilized a set of 22 explosive keywords (e.g. anfo, c-4, etc.) provided by domain experts and compared the results of submitting keyword-based, pattern-based and spice-expanded queries. For the keyword-based search, 22 queries were submitted and the top 20 results were retrieved for each (i.e. 440 URLs). For the pattern-based search, we mapped the keywords to the HME patterns. Specifically 396 (22 keywords × 18 explosive patterns, Sect. 5.2) queries were submitted and the top 20 URLs for each mapped keyword were retrieved (i.e. 440 URLs) after merging the results using a linear ranking system. Finally, for the spice-expanded queries, we used the 2 best performing keyword spices during the training process (Sect. 5.2) and submitted 44 queries (22 keywords × 2 keyword spices) retrieving the top 20 results (i.e. 440 URLs) for each expanded keyword after merging the results using a linear ranking system. In total, 1320 URLs were retrieved, resulting in 720 unique URLs (after duplicate elimination), which were manually annotated. Table 3 compares the results of the three approaches.

Table 3. Querying evaluation results

	Keywords	Keyword + Patterns	Keyword + K.S.
MAP	0.70	0.87	0.71
Relevant URLs	232	233	97
Newly discovered relevant URLs	–	161	48

The improvement of the ranking results when the pattern-based approach is employed as opposed to the simple keyword-based approach indicates the significance and the usefulness of the pattern-based approach for the HME domain. Also, the “keyword spices” methodology, slightly improves the ranking when compared to the keyword-based approach. Furthermore, retrieving high recall results is of equal importance. Employing both the pattern-based and the “keyword spices” approach results in a significant increase in the number of the discovered relevant results compared to the simple keyword-based approach, which means that the recall is increased.

8 Conclusions

This work proposed an interactive search engine that was adapted in the context of the discovery and retrieval of Web resources containing HME information. It is envisaged that within such an application, which is driven by the distinctive nature of its user requirements, the tools developed provide LEAs with the technology they require to acquire more knowledge on HMEs in order to tackle the threat they pose. The application has indicated the significant adaptability of the domain-specific search approaches to the HME domain. Future work includes more extensive evaluation of the interactive search engine in terms of its usability, effectiveness, and efficiency by LEA personnel, and domain experts, in large-scale user studies that will take place.

Acknowledgement. This work was supported by the HOMER (312388) FP7 project.

References

1. Olston, C., Najork, M.: Web crawling. J. Found. Trends Inf. Retrieval **4**(3), 175–246 (2010)
2. Agarwal, G., Kabra, G., Chang, K.C.C.: Towards rich query interpretation: walking back and forth for mining query templates. In: 19th ACM International Conference on World Wide Web (WWW 2010), pp. 1–10 (2010)
3. Oyama, S., Kokubo, T., Ishida, T., Yamada, T., Kitamura, Y.: Keyword spices: a new method for building domain-specific web search engines. In: 17th International Joint Conferences on Artificial Intelligence, IJCAI-2001, pp. 1457–1463 (2001)
4. Oyama, S., Kokubo, T., Ishida, T.: Domain-specific web search with keyword spices. J. IEEE Trans. Knowl. Data Eng. **16**(1), 17–27 (2004)
5. Stenersen, A.: The internet: a virtual training camp? J. Terrorism Polit. Violence **20**, 215–233 (2008)
6. Chen, H.: Dark web: exploring and mining the dark side of the web. In: Domenach, F., Ignatov, D.I., Poelmans, J. (eds.) ICFCA 2012. LNCS, vol. 7278, p. 1. Springer, Heidelberg (2012)
7. Fu, T., Abbasi, A., Chen, H.: A focused crawler for Dark Web forums. J. Am. Soc. Inf. Sci. Technol. **61**(6), 1213–1231 (2010)
8. Chen, H., Chung, W., Qin, J., Reid, E., Sageman, M., Weimann, G.: Uncovering the dark Web: a case study of Jihad on the Web. J. Am. Soc. Inf. Sci. Technol. **59**(8), 1347–1359 (2008)
9. Kalpakis, G., Tsikrika, T., Markatopoulou, F., Pittaras, N., Vrochidis, S., Mezaris, V., Patras, I., Kompatsiaris, I.: Concept detection on multimedia web resources about home made explosives. In: 10th International Conference on Availability, Reliability and Security (ARES 2015), pp. 632–641 (2015)
10. Tsikrika, T., Kalpakis, G., Vrochidis, S., Kompatsiaris, I., Paraskakis, I., Kavasidis, I., Middleton, J., Williamson, U.: A framework for the discovery, analysis, and retrieval of multimedia homemade explosives information on the Web. In: 10th International Conference on Availability, Reliability and Security (ARES 2015), pp. 601–610 (2015)

11. Tsikrika, T., Moumtzidou, A., Vrochidis, S., Kompatsiaris, I.: Focussed crawling of environmental web resources: a pilot study on the combination of multimedia evidence. In: 1st International Workshop on Environmental Multimedia Retrieval (EMR 2014), in conjunction with the ACM Conference on Multimedia Retrieval (ICMR 2014), pp. 61–68 (2014)
12. Pant, G., Srinivasan, P.: Learning to crawl: comparing classification schemes. ACM Trans. Inf. Syst. **23**(4), 430–462 (2005)
13. Quinlan, J.R.: C4.5: Programs for Machine Learning. Elsevier, Amsterdam (1994)
14. Cortes, C., Vapnik, V.: Support-vector networks. J. Mach. Learn. **20**(3), 273–297 (1997)