A Framework for the Discovery, Analysis, and Retrieval of Multimedia Homemade Explosives Information on the Web

Theodora Tsikrika, George Kalpakis, Stefanos Vrochidis, Ioannis Kompatsiaris
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
{theodora.tsikrika, kalpakis, stefanos, ikom}@iti.gr

Iraklis Paraskakis, Isaak Kavasidis
Information & Knowledge Management Research Cluster
South East European Research Centre
The University of Sheffield, International Faculty,
CITY College
Thessaloniki, Greece
{iparaskakis, ikavasidis}@seerc.org

Jonathan Middleton, Una Williamson
Police Service Northern Ireland
Belfast, UK
{Jonathan.Middleton, Una.Williamson}@psni.pnn.police.uk

***Abstract*—This work proposes a novel framework that integrates diverse state-of-the-art technologies for the discovery, analysis, retrieval, and recommendation of heterogeneous Web resources containing multimedia information about homemade explosives (HMEs), with particular focus on HME recipe information. The framework corresponds to a knowledge management platform that enables the interaction with HME information, and consists of three major components: (i) a discovery component that allows for the identification of HME resources on the Web, (ii) a content-based multimedia analysis component that detects HME-related concepts in multimedia content, and (iii) an indexing, retrieval, and recommendation component that processes the available HME information to enable its (semantic) search and provision of similar information. The proposed framework is being developed in a user-driven manner, based on the requirements of law enforcement and security agencies personnel, as well as HME domain experts. In addition, its development is guided by the characteristics of HME Web resources, as these have been observed in an empirical study conducted by HME domain experts. Overall, this framework is envisaged to increase the operational effectiveness and efficiency of law enforcement and security agencies in their quest to keep the citizen safe.**

Keywords-homemade explosives; multimedia; discovery; analysis; retrieval; recommendation; knowledge management

I. INTRODUCTION

As the global economy attempts to recover, there are increasing pressures on public and government authorities to improve the efficiency of their organisations, while also increasing their effectiveness. This is equally the case with law enforcement and security agencies that, whilst charged with upholding the rule of law and keeping the citizen safe, need to consider innovative ways of meeting these challenges. Indeed as law enforcement agencies continue to face increasing demands on their resources in order to deliver a policing service and reduce crime, they need to consider alternative opportunities to meet these demands. Significantly the development of technology has already provided solutions to a number of the challenges being faced, thus releasing personnel to undertake other duties without increased demand on the organisation. The continuing evolution in technology provides opportunities for law enforcement and security agencies to face current and evolving threats without over-burdening resources.

One challenge being realised and continuing to evolve is the proliferation of the use of the Web and social media platforms, which is resulting in a significant increase in information sharing globally. This growth has facilitated the diffusion of knowledge and experience in many different ways, some of which present a threat to society, both on a national and international level, thus presenting challenges for both law enforcement and security agencies, charged with investigating and protecting national and international interests.

The information being shared online includes material on supporting and undertaking acts of terrorism at varying levels, exploitation, financial crime, and cybercrime. In some cases, even the innocent uploading of information can also be exploited for such activities; this includes information on the manufacture and use of homemade explosives. The perception that this domain contains information exclusive to the terrorist or criminal is not accurate, as there has been indeed an increasing trend through Web and social media sites for individuals, academic institutes, and media organisations to share multimedia content on hobbies or research they have undertaken with many different materials, to determine or demonstrate their ability to be used as precursors in homemade explosives [1].

The challenge for society through law enforcement and security agencies is to better understand and counter the ability for subversive use of this information. Whether it is for the use in an act of terrorism, serious and organised crime, or an individual with particular reasons to instigate an IED (Improvised Explosive Device) attack on a target, there is a challenge to develop technology, operational awareness, and tactical advantage in order to counter this potential threat. The proliferation of information provides the subversive with the ability to easily research and understand the construction of IEDs and the manufacture of homemade explosives using everyday household goods, easy to purchase items, or materials which, although

978-1-4673-6590-1/15 $31.00 © 2015 IEEE
DOI 10.1109/ARES.2015.86

controlled, can still be easily acquired. Therefore, law enforcement and security agencies need technologies that would allow them to tackle this threat through the automatic identification of such HME information and its understanding so as to enable interested stakeholders to access it in an effective and efficient manner.

In order to meet this challenge, this work proposes a novel framework for the discovery, analysis, and retrieval of multimedia content on homemade explosives (HMEs) on the Web, with particular focus on HME recipe information. Such information can be found on various types of Web resources, such as Web pages, blogs, forums, and social media posts (e.g. on social networking sites, such as Facebook[1] and Twitter[2], and social content sharing platforms, such as YouTube [3] and Flickr [4]), that may express their content in multiple languages and in various modalities, such as text, images, and video. The proposed framework is being developed in the context of the activities of the HOMER [5] (HOmeMade Explosives and Recipes characterisation) the EU FP7 project and corresponds to a **Knowledge Management Platform** that enables the interaction with knowledge about HMEs through a user friendly intuitive interface, while it also supports and promotes the collaborative sharing and exchange of such information, and consists of three major components that provide several advanced functionalities:

1. an **HME Discovery** component that enables the identification of Web resources containing HME information; once discovered, such resources are uploaded to the Knowledge Management Platform (following a validation process),
2. a **Content-Based Multimedia Analysis** component that detects HME-related concepts on multimedia content so as to support their retrieval,
3. an **Indexing, Retrieval, and Recommendation** component that processes the available HME information to enable its (semantic) search and provision of similar information.

The main contribution of this work is the integration of advanced state-of-the-art technologies in a novel framework for the discovery, analysis, and retrieval of heterogeneous Web resources containing multimedia HME information, that has been developed in a user-driven manner in close collaboration with and based on the requirements of law enforcement agencies and security personnel, as well as other interested stakeholders, such as HME domain experts. To the best our knowledge, this is the first attempt to develop a holistic framework tailored to the discovery, analysis, retrieval, and recommendation of multimedia HME information.

The remainder of this paper is structured as follows. Section II discusses the user requirements that motivate and underpin the development of the proposed framework. Section III reviews related work. Section IV presents the architecture of the framework, while Sections V, VI, and VII describe its major components: (i) discovery of HME Web resources, (ii) content-based multimedia analysis, and (iii) indexing, retrieval, and recommendation, respectively. Finally, Section VIII concludes this work.

[1] http://www.facebook.com
[2] http://www.twitter.com
[3] http://www.youtube.com
[4] http://www.flickr.com
[5] http://homer-project.eu/

II. End-User and Data Requirements

This section discusses the requirements of the end users of the proposed framework through appropriate use cases, each described by a scenario and the envisaged application of the framework in this scenario. These use cases have been provided by interested stakeholders, in particular, by law enforcement and security agents, as well as HME domain experts, participating in HOMER, who have offered guidance for a user-oriented development. In addition, the main findings of an empirical study conducted for identifying the types and characteristic features of HME Web resources are also presented, so as to further identify, from a data perspective, the requirements for the discovery, analysis, retrieval, and recommendation components of the proposed framework.

A. User Perspective

Use Case 1. **HME incident investigation**. *Scenario*: Law enforcement and security agencies need to be able to identify quickly the perpetrators of a bomb threat or attack. In the case an HME device is found, e.g. in a public place, and deactivated by an EOD (Explosive Ordnance Disposal) team, the device is further analysed in police laboratories for identifying the chemical ingredients it contains and the way it was constructed[6]. *Application*: Once this information becomes available, the proposed framework can support several searches within its knowledge management platform for finding HME recipes containing some (or all) of these substances, their chemical characteristics and dangerousness, as well as incidents where similar substances and/or devices have been used. In addition, it offers recommendations corresponding to content that may not contain the query terms, but it is considered complementary to what the user is seeking. The results of such searches and recommendations have the potential to lead to the identification of groups or individuals with the same or similar modus operandi.

Use Case 2. **Intelligence gathering**. *Scenario*: Law enforcement and security agencies, as well as HME experts, need to be able to monitor the advancements in the manufacture of HMEs and such information being shared online. Consider, for instance, the case that there is intelligence that a specific substance not previously used in an incident, and thus not available on the platform, is currently being considered for subversive use. *Application*: In such cases, the proposed framework can support the discovery of information about this substance on the Web and social media. To this end, the framework offers search functionalities that support the users' query formulation in a semi- or fully automatic mode, and also provides the crawling of Web sites, forums, and social networks for identifying content focussed on the information being sought.

Use Case 3. **Multimedia analysis**. *Scenario*: Law enforcement and security agencies are interested in the automatic analysis of multimedia content available on the

[6] This may also take place in case the bomb has exploded; then, its remains are examined.

Web and in particular on social content sharing platforms, such as YouTube. Their goal is both to identify videos containing HME-related content among all available information, and also to detect specific HME-related objects within such videos for extracting useful information. For instance, the automatic identification of video segments showing the particular chemical substances being used for the manufacture of an HME, or the 'fireball' from various blasts (e.g. in extremist videos) allows them to obtain evidence that could be useful in future investigations. *Application*: To this end, the proposed framework offers the detection of HME-related concepts (i.e. objects) in visual content, such as images and videos embedded in Web resources.

B. HME Web Resources: Types and Characteristics

To determine the types and characteristics of HME Web resources, four of the HME domain experts who participate in the HOMER project conducted an empirical study through manual examination and observation of HME information they manually discovered on the Web and social media. In particular, a dataset consisting of 129 HME Web resources was thoroughly studied in a recursive manner based on the well-established empirical cycle methodology [2], so as to identify the most prominent types of HME Web resources, the meaningful HME information they convey and the way this information is encoded. Given the availability of such content in multiple languages, two languages of particular significance to interested stakeholders due to their prominence in the global landscape were selected: English and Arabic. Thus, the constructed dataset consists of 76 HME Web resources in English and 53 in Arabic.

The analysis indicated that there are several different types of Web resources containing HME information, including typical Web pages (composed of static or/and scripted content), blogs, torrent files, forums, electronic books, videos on social content sharing platforms, and posts on social networking sites. In addition, such HME Web resources are available both on the Surface Web (i.e. they correspond to the free and publicly available Web resources that are typically indexed by Web search engines), and also on the Hidden Web [3], including resources available through anonymous networks, such as the Tor[7] anonymity network which is based on the onion routing system [4]. Moreover, these HME Web resources contain rich sets of recipes expressed in multiple modalities and covering a wide and diverse range of HMEs.

This analysis further indicated that the most meaningful HME recipe information conveyed on all these different types of Web resources corresponds to the ingredients used for the synthesis of the HME, followed by their measurements (i.e. the required quantities) and the procedure to follow for the HME synthesis. Additional features that were occasionally observed, to different degrees across the different types of HME Web resources and the different languages, include the equipment needed for the synthesis process, the chemical formulas representing the ingredients, possibly their reactions, and the resulting explosives, and chemical properties of the synthesised HMEs, such as their melting point. This meaningful HME recipe information is typically encoded as text, but may also be additionally encoded in the multimedia content (e.g. images and videos) occasionally embedded in Web resources. Their title, and to a lesser extent their metadata and URL, are informative, while the comments posted on HME Web resources, and particularly those in forums and social content sharing platforms, may also include relevant material. Moreover, crawling strategies could exploit the observation that the anchor text of the hyperlinks leading to relevant information often contains HME-related terms and that HME Web resources, particularly those in English, are often hosted on Web sites that also contain additional relevant material.

C. Requirements

All three use cases reflect the challenge of having to deal with large amounts of data and information in an effective and efficient manner. However, there are also very significant differences in the required functionalities, with each having a different focus on the ultimate goal and the operations being performed to reach that goal. For example, police officers are mainly interested in finding information about particular HMEs and incidents where such HMEs have been deployed, while intelligence officers and domain experts aim to gather and analyse information mainly for monitoring purposes.

In addition, the observation that such HME information is hosted in many different forms on the Web requires that the framework is capable of handling heterogeneous data. Moreover, given that this information is continually fluid, potentially moving between various parts of the Web or being removed when no longer required or suspecting of being identified, requires that the framework provides a means for automated discovery, analysis, and retrieval of this information, with minimal human interaction.

The borderless nature of the Web also indicates that there is a requirement for such tools to consider and demonstrate an ability to undertake their role in a multi-lingual environment. To this end, the framework is built on the basis of language-independent components (to the extent possible); here, this work demonstrates it for the English language, but additional languages, such as Arabic, are currently being incorporated. Furthermore, to ensure that as much information can be analysed and retrieved for the end user, the multimedia content needs to be analysed.

Finally, law enforcement and security agencies also need to understand the validity of the information and its ability to provide sufficient information for subversive use. It could be reasonably assumed that there will be limitations on technology to automatically determine the validity of HME information, as the discovered information may also pertain to completely innocent causes. For example, the term sugar could be used as a precursor, yet it is also used innocently every day in food produce. With this in mind, there is currently still a need to use human interaction for validating the information being automatically discovered.

Overall, the functionality of the proposed framework considers all the requirements imposed by the described use cases and the characteristics of HME Web resources.

[7] https://www.torproject.org

III. RELATED WORK

Scientific research towards the study of extremist (terrorist-related) content on the Web [1] has paved the way for the development of techniques and tools that aim to collect and analyse Web content generated by international terrorist groups, including websites, forums, blogs, social networking sites, videos, etc. To this end, the most comprehensive suite of multilingual data mining, text mining, and Web mining tools for performing link, content, sentiment, authorship, and video analyses has been developed in the context of the Dark Web project at the University of Arizona [5]. However, this project has addressed the whole breadth of terrorist and extremist content, rather than HME information, as done here. Close to this work is also the system developed in the context of the activities of the EU FP7 Security Research Project CAPER [8] (Collaborative information, Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime). Their system was also created in cooperation with European law enforcement agencies and aims to build a common collaborative and information sharing platform for the detection and prevention of organised crime. Again, they deal with a wider scope of information, rather than only HMEs, and they also put particular emphasis on the analysis of social networks [6], mainly Facebook, rather than all different types of available Web resources. Our work thus proposes a novel integrated framework that focusses on HME information and which is being developed in a user-driven manner with the goal to fulfill the requirements of its end users.

IV. FRAMEWORK ARCHITECTURE

This section provides a high level overview of the architecture of the developed framework. As shown in Figure 1, the framework corresponds to a **knowledge management platform (KMP)** that enables the various stakeholders (e.g. police and anti-terrorism/intelligence officers, as well as chemists and HME experts) to interact with knowledge about HMEs, while it also provides several advanced functionalities implemented by the following components: (i) the HME discovery component, (ii) the HME analysis component, and (iii) the HME indexing, retrieval, and recommendation component.

Given the sensitivity of the information handled, it is imperative that access to it is controlled, as well as that the information goes through a validation process. To that effect, the KMP has adopted a role-based access system. The permissions that can be granted relate not only to content access, but also define whether a user can perform an action or not (e.g. upload/validate content, perform semantic search, initiate HME discovery, etc.). Users may also belong to different groups, thus access to specific resources and permissions for specific actions can be limited to users belonging to specific groups.

Content creation is done with the use of light intuitive Web 2.0 tools such as *blogs* and/or *wikis*. Users can decide to upload their content as a blog whenever they want the content to be editable only by themselves, while a wiki permits all the users of a group to edit it. The users must tag the content in order to enable intelligent processing of it during retrieval. The tagging process is done primarily automatically, but the user has the final say by either adding and/or removing tags. Regarding the platform's advanced functionalities that aim to aid the users in finding what they are searching for, the **HME discovery** functionality enables the users to perform customised searches on the Web and social media, by employing public general-purpose Web search engines and focussed crawlers. The former relies both on semi-automatic query formulation approaches based on query patterns and also on automatic query expansion methods based on the generation of "keyword spices". The focussed crawlers instead start from a predefined set of seed Web resources and subsequently traverse the Web link structure with the goal to identify Web resources having HME-relevant content. In both cases, the results of the discovery process go through a classification step for reducing the potential noise (see Section V).

After the HME Web content has been discovered, **multimedia content-based analysis** is performed aiming at processing the multimedia content and associate low-level visual features with high-level HME concepts so as to support their retrieval (see Section VI). The HME content that has been discovered and analysed is then uploaded to the knowledge management platform following an extra validation step (by a user authorised to perform such operations) that determines its validity and possibly its dangerousness and other characteristics.

Once the HME content has been discovered and its multimedia parts have been analysed, it goes through an **indexing** process so as to support its **retrieval** through a semantic search functionality that allows users to discover hidden semantic relationships between concepts included in the content, which cannot be found if a simple term-by-term search is performed; this is achieved by exploiting DBpedia[9] relations. In addition, the platform also features a variety of **recommendation** systems that offer to users information complementary to what they are interested in. These recommendations result from text processing algorithms and are based on how similar the currently viewed content to other content is. The similarity metrics exploited currently in the platform are the content, the tags and their semantic relationships, and the browsing history (see Section VII).

Next, the major components of the framework are described.

V. DISCOVERY OF HME WEB RESOURCES

The **discovery** of HME Web resources is addressed as a domain-specific search problem [7] and is based on a hybrid infrastructure (see Figure 1) that combines into a joint workflow two different approaches: (i) a W eb crawler focussed on the HME domain that, 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, and (ii) the submission of HME domain-specific queries to general-purpose search engines; such queries are either formulated using query patterns in conjunction with domain knowledge, or are expanded based on the methodology of "keyword spices" (i.e. domain-specific keywords generated with the aid of

[8] http://www.fp7-caper.eu

[9] http://wiki.dbpedia.org/

supervised machine learning techniques) [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 performing more focussed crawls on the topic. Moreover, the application of focussed crawling avoids dependencies on external services (i.e. existing search engines), thus ensuring the long-term viability of the framework.

Then, the discovered Web resources pass through a series of pre-processing steps (not depicted in Figure 1). First, they pass through a *language identification* component, which is a prerequisite for performing text-based analysis; to this end the JLangDetect[10] library is employed. This process is followed by two optional

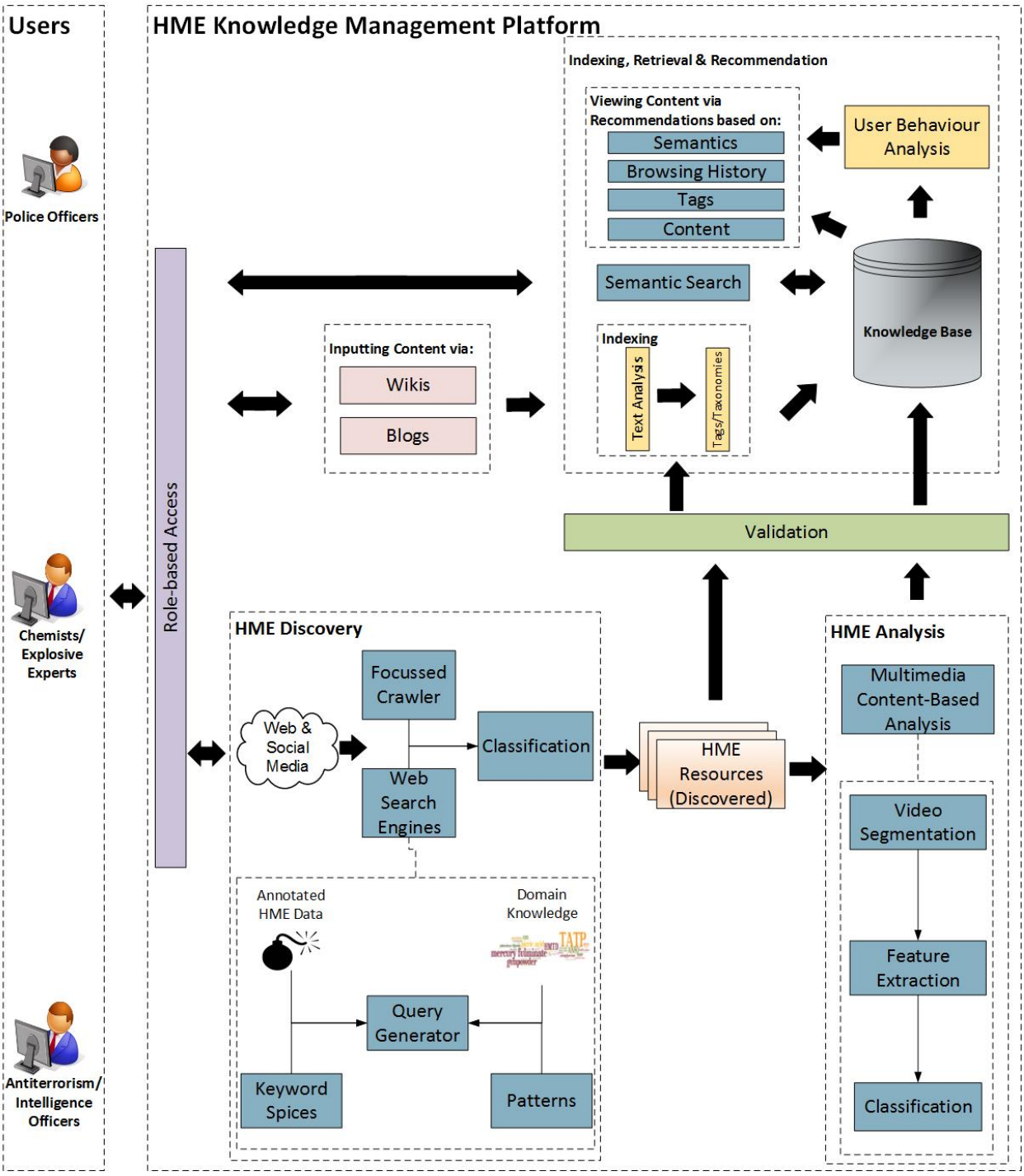

Figure 1: Architecture of the framework for the discovery, analysis, indexing, retrieval, and recommendation of HME information

[10] https://github.com/melix/jlangdetect

components, *content extraction* and *duplicate detection*. The first component is realised by employing the boilerpipe algorithm[11] [9] that separates the actual textual content of a Web resource from the entire existing surplus "clutter", such as navigational elements, templates, advertisements and so on. That kind of "noise" may deteriorate the accuracy of the analysis and thus needs to be detected and removed properly. The second process aims at detecting Web resources having identical or highly similar (i.e. near-duplicate) content. This is a common phenomenon observed on the Web and, if not addressed properly, it might result in significant redundancy within the final set of discovered HME Web resources. To this end, a plagiarism detection software[12] [10] is currently being used, but alternative methods proposed for Web page similarity detection [11] could also be employed.

Finally, post-processing filtering is performed based on a classification process for reducing the potential noise in the results obtained by both discovery approaches.

Next, the major modules of the HME discovery component are presented.

1) Focussed Crawling

Starting from a set of predefined (seed) Web pages on a topic, *focussed* (or *topical*) *crawlers* fetch (i.e. download) these pages, mine their content to extract the hyperlinks they contain, and select the ones that would lead them to other pages relevant to the topic. This process is iteratively repeated until the desired number of pages is fetched or some other stopping criterion is applied. To predict the benefit of fetching an unvisited Web page is a challenging task, since its relevance to the topic at hand needs to be estimated based solely on evidence obtained from the already downloaded pages. To this end, state-of-the-art approaches [12] adopt classifier-guided crawling strategies based on supervised machine learning methods that rely on two sources of evidence for classifying the hyperlinks: (i) their local context, i.e. the textual content appearing in the vicinity of the hyperlink in the parent page, such as its anchor text and its surrounding text, and/or (ii) global evidence associated with the entire parent page, such as its textual content and its hyperlink structure.

Our platform employs a classifier-guided focussed crawling approach for the discovery of HME Web resources. To this end, it estimates the relevance of a hyperlink to an unvisited resource based on its local context. Motivated by the results of the aforementioned empirical study that indicated 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 and HME-related keywords), we follow recent research [13] 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[13] 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, normalised 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 [14]. 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 focussed crawler is based on Apache Nutch[14] (version 1.9), a robust open-source Web crawling and search software project written in Java. It runs on top of Hadoop[15], which allows for the distributed processing of large datasets across clusters of computers using simple programming models. It is designed to scale-up from single servers to thousands of machines, each offering local computation and storage. In addition to its scalability, it is highly extensible and configurable.

Nutch can run on several protocols (e.g. http(s), file, ftp, etc.), parse different formats (e.g. html, pdf, doc, etc.), and extract all included metadata. Its fetcher functions in a multi-threaded fashion, where each thread follows by default a strict politeness policy (which can be configured to a more lenient mode) for avoiding overloading Web servers and sites. However, its core functionality does support focussed crawling and thus code customisation was performed for supporting the methods described above. In particular, the URL extraction policy of its parser was tailored to our requirements. Additional customisations were also applied for handling some of the types of HME Web resources that need special treatment, such as password-protected ("private") Web resources, and Web resources with limited access due to restrictions enforced by the robots exclusion protocol rules. In both cases, such Web resources are identified and appropriately tagged, so that human intervention can be sought.

Moreover, the focussed crawler has been configured so that it can be set to traverse the Tor network, in order to discover HME-related *.onion* pages hosted on Tor Hidden Services. To this end, the Tor facility is enabled on the machine running the crawler and the crawler takes advantage of an HTTP to SOCKS proxy server infrastructure, which constitutes the mediating point between the crawler and the Tor network. Finally, the available parameter list has been extended so as to add the newly introduced variables, such as the score threshold that determines the necessary relevance value that a URL needs to exceed in order to be accepted by the classifier.

2) Querying Search Engines

This component submits domain-specific queries to existing general-purpose search engines. Currently, the Yahoo! BOSS API[16] is employed, but it can be easily replaced by other APIs (provided for instance by Google,

[11] http://code.google.com/p/boilerpipe/

[12] http://rp-www.cs.usyd.edu.au/~scilect/sherlock/

[13] This limit of x characters is automatically extended so as to guarantee that it will not split any of the words lying at the edges of the window.

[14] http://nutch.apache.org/

[15] http://hadoop.apache.org/

[16] https://developer.yahoo.com/boss/search/

Bing, etc.). The domain-specific queries are generated in a *manual*, *semi-automatic*, or fully *automatic* fashion.

The **manual** approach assumes the actual involvement of domain experts and stakeholders from law enforcement and security agencies who formulate queries based on their knowledge and expertise. Their most important advantage is their profound knowledge of the core characteristics and terminology of the HME domain, stemming from their extensive experience and substantial involvement in the field. However, in order to obtain effective search results, it is also very important for an expert to have an adequate understanding of the rules that govern the way a query should be formed and expressed. Given the inherent difficulty of this task, improvements in the effectiveness of such searches are typically achieved either through filtering or/and post-retrieval analysis [15], or through the use of (semi-)automatic query formulation methods.

The **semi-automatic** approach requires the availability of an initial set of seed queries (e.g. obtained from query logs) 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 [16]. Here, an initial set of 64 queries (45 in English and 19 in Arabic) that was formulated by domain experts and law enforcement agencies personnel and was used for the successful discovery of HME Web resources through general-purpose search engines, is used as the seed set of queries; examples range from simple queries, e.g. "tatp" to more verbose ones, e.g. "how to make mercury fulminate at home". Once the keywords appearing in all queries are mapped to discrete concepts (e.g. the keywords "preparation" and "acetone peroxide" are mapped to the concepts "*action*" and "*explosive*"), 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 acetone peroxide" becomes "*action explosive*". This results in producing a set of discrete patterns that can be used for automatic query generation. For example, the pattern consisting of the concepts "*action explosive*", where "*action*" corresponds to keywords such "how to make", "preparation", "synthesis" etc., may be instantiated to several different queries for each of the explosives of interest. Therefore, once an end user expresses interest in discovering HME Web resources about a given explosive, all query patterns containing the concept "*explosive*" will be automatically instantiated for that particular explosive, will be submitted in parallel to the search engine, and the results of all these queries will be merged before being presented to the user. This methodology is largely language-independent, with the exception of the initial steps that requires language-specific resources or human experts for mapping the extracted keywords to concepts.

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) expressions (referred to as "keyword spices") [8] that aim to characterise 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 following methodology is applied. First, a set of Web resources annotated by domain experts with respect to their relevance to the HME domain is split into two disjoint subsets: (i) the training set for generating the initial keyword spices, and (ii) the validation set for simplifying them. Then, the nouns found in these Web resources[17] are extracted so as to be used as domain-specific keywords. Based on the training set, a (binary) decision tree is constructed using the information gain measure without any pruning techniques and a decision tree learning algorithm is applied for discovering the keyword spices. To this end, the C4.5 algorithm [17], and specifically its J48 implementation in the Weka[18] machine learning tool, is used. The internal nodes of the decision tree correspond to the extracted keywords and its leaves to class labels. This results in a decision tree that can be expressed as a set of rules or as a Boolean disjunctive normal form; these are the initial keyword spices.

Similar to rule post-pruning, these initial keyword spices are simplified by iteratively removing keywords (or entire conjunctions) if their removal increases the F-measure (i.e. the harmonic mean of precision and recall) over the validation set, i.e. if their removal improves the effectiveness compared to them occurring in the query. This process is repeated until there is no keyword (or conjunction) that can be removed without decreasing the F-measure. For example, in a set of 1157 Web resources (517 relevant to the HME domain and 640 non-relevant), the initial keyword spice corresponds to the following complex Boolean expression (where ^ signifies the NOT operator): *problem OR bombs OR home OR time OR impact OR ^glass OR heating OR terms OR acid OR ^power OR ^rights OR ^time OR grams OR alcohol OR cap OR fuel OR reaction OR (explosive AND ^petn) OR (explosive AND ^world) OR (explosive AND acid)*, whereas the simplification leads to the simpler expression: *heating OR grams OR cap OR fuel OR reaction OR (explosive AND acid)*.

This is a largely language-independent approach, with the only language-dependent step being the one that performs textual feature extraction for identifying the domain-specific keywords to be used as nodes in the decision tree. Typically, this step applies tokenisation, stopwords removal, and possibly lemmatisation [19]; stemming is not usually applied as the goal is to obtain keywords that could be used for query expansion.

3) Post-processing Classification

The resources discovered through focussed 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, tokenisation, stopwords removal and stemming are applied, and its textual feature vector is

[17] Here, the Part-Of-Speech (POS) tagger developed by the Stanford Natural Language Processing (NLP) (group http://nlp.stanford.edu/software/tagger.shtml) is used.

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

[19] http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

generated using the *tf.idf* term weighting scheme, where *tf(t,d)* is the frequency of term *t* in sample *d*, normalised 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 classifier is built using an RBF kernel, while 10-fold cross-validation is performed for selecting the class weight parameters. This classifier is implemented using the libraries of the Weka machine learning software.

The above process includes several language-dependent steps, namely tokenisation, stopwords removal, and stemming, whereas the rest are based on statistical properties of the text and are thus language-independent. For the English language, these language-dependent steps are implemented as follows. Tokenisation is performed using a Java implementation of a tokeniser based on regular expressions, stopwords removal is performed based on a list of 636 common English stopwords[20], while stemming is performed using Porter's algorithm for the English language [18].

VI. Content-based Multimedia Analysis

The multimedia analysis component aims to associate the low-level visual features of multimedia content (i.e. images and videos) embedded in Web resources with high-level HME-related semantic concepts, with the ultimate goal to detect HME-related objects on visual content so as to support retrieval in response to end-user queries. To this end, a state-of-the-art multimedia concept detection component for the automatic classification of visual content to a set of predefined semantic concepts is employed and consists of the following sequential steps:

1. **video decoding** for extracting the most representative frames from videos for further processing,
2. **feature extraction and representation** for extracting a set of descriptors that effectively characterise the visual content, and
3. **classification** for training and subsequently applying the trained models so as to classify the multimedia content to a set of predefined HME-related semantic concepts.

Regarding the **video decoding**, the segmentation of a video into shots is performed using a variation of the algorithm introduced in [19]. The detection of shot boundaries is based on the assessment of the visual similarity between neighboring frames of the video. For this, the visual content of each frame is represented by computing a set of local (i.e. the ORB descriptors proposed in [20]) and global (i.e. HSV color histograms) descriptors, thus allowing the detailed matching of a pair of frames and the effective detection of their differences both in color distribution and at a more fine-grained structural level. Then, shot transitions are detected by quantifying the visual similarity between successive or neighboring frames of the video, and comparing it against experimentally specified thresholds that indicate the existence of abrupt and gradual shot transitions. Finally, the overall detection efficiency of the algorithm is enhanced by using dedicated detectors for the identification of dissolves and wipes, and for filtering false alarms.

The **feature extraction** step exploits local features based both on the popular SIFT descriptor [21] along with its colour-based variations (i.e. RGB-SIFT and opponent-SIFT) [22], and on the widely-used and efficient SURF descriptor [23] and its recently introduced colour-based variations (i.e. RGB-SURF and opponent-SURF) [24]. Both SIFT and SURF perform scale-invariant feature extraction. SIFT is capable of reliably detecting objects, even among clutter and under partial occlusion. SURF is popular for its high performance and it is well-known for its robustness against several different image transformations. Feature encoding is realised using VLAD [25], which on one hand is capable of reliably representing an image by aggregating local descriptors in vector space, and on the other hand allows for remarkable dimensionality reduction by performing principal component analysis (PCA), without impacting its accuracy. The approach adopted is described in detail in [24]. The obtained VLAD vectors serve as input to the classification step.

The **classification** is based on logistic regression where a linear model is trained using logistic regression for learning associations between visual characteristics and semantic concepts, based on a set of annotated images/video frames serving as training data (where each image/video frame is characterised either as relevant to a higher-level concept or not). The trained model is able to classify an unlabelled image/video frame by estimating its confidence score of being relevant to the concept under consideration. For each descriptor (i.e. SIFT and SURF descriptors, and their variations), a distinct classifier is trained for each concept and an average prediction score per concept is produced, taking into account the confidence scores calculated by each classifier.

Given the lack of available concept lexicons for the HME domain, a set of HME-related concepts was determined based on end-user requirements and also on the findings of the aforementioned empirical study. This resulted in a list of 15 HME-related concepts, including concepts related to the appearance of HME material (e.g. "*powder*", "*granules*", etc.), equipment used for the synthesis of HMEs (e.g. "*glassware*"), explosions resulting from using HMEs (e.g. "*fire*", "*smoke*", etc.), and improvised explosive devices containing HMEs (e.g. "*device*"). For each of these concepts, a classifier has been trained and is available for annotating unlabelled multimedia content so as to support its retrieval.

VII. Indexing, Retrieval, and Recommendation

Given the intrinsic properties of the HME domain, fast and accurate access to data is of critical importance. In fact, the knowledge management platform was designed with efficiency and accuracy in mind, in order to offer to law enforcement agents the ability to find quickly not only the requested information, but also similar information that may be useful in the situation they find themselves in. For this reason, the platform was implemented by using the most efficient algorithms for indexing and retrieval.

In particular, the **indexing** of the resources in the platform is implemented by an NLP pipeline based on the

[20] https://code.google.com/p/stop-words/

OpenNLP[21], which is composed of standard methods for obtaining the keywords of documents (i.e. tokenisation, stopwords removal, and stemming) combined to a Vector Space Model text representation [3] On the basis of these representations, an inverted index [25] is created so as to render the retrieval of the resources more efficient. While these methods are straightforward, they are very stable, accurate, and fast, thus satisfying all user requirements.

The indexing is also complemented by four different taxonomies, and in particular, folksonomies (i.e. user maintained taxonomies). The three of them refer to specific aspects of HMEs or cases involving HMEs, while the last one is a general-purpose folksonomy. The three HME-specific folksonomies relate to:

1. **precursors,** which contains terms regarding chemical substances and materials,
2. **investigation,** which contains terms regarding an event or case involving the use of HMEs (e.g. place names, perpetrator names etc.), and
3. **construction,** which contains terms regarding the construction or methods of construction of HMEs (e.g. equipment used, non-explosive materials, casings, etc.)

The general-purpose folksonomy contains all the terms that are unrelated to the aforementioned folksonomies.

The **retrieval** of the information consists of returning a set of resources in response to a keyword-based query. To this end, the platform implements two different search mechanisms:

1. **Simple search:** In this case, the query terms are compared against the inverted index obtained during the indexing phase and the resources that contain the most occurrences of the query terms are returned in a descending order.
2. **Semantic search:** In high-risk situations, one cannot expect from users to be able to identify the precise terms that they would need to use so as to obtain effective retrieval results. To deal with this problem, the semantic search of the CLIIP framework [27] is employed. For example, let's suppose that a police officer conducts an investigation about a case where a pipette was used during the construction of an HME. The CLIIP framework finds the general categories of the "pipette", i.e. "Laboratory glassware", "Laboratory equipment", "Microbiology equipment", and "Volumetric instruments", and returns to the user all related content that contains also terms from these categories, e.g. terms such as "graduated cylinder", "blender", "dripstick" etc.

Moreover, several recommender systems are deployed aiming to provide the officers with relevant information. The recommenders are based on similar characteristics that the content may have with the other resources in the platform. To this end, the following five types of recommender systems are deployed:

1. **Tag recommendation:** Whenever a user submits a new resource in the system the textual content is automatically processed by online services and NLP algorithms in order to identify the most prominent concepts and extract them as tags. Finding the most important concepts in text is not always a trivial task. The online services used for this task are: (i) Yahoo Term Extractor[22], (ii) Opain Calais[23], (iii) Zemanta[24] and (iv) DBpedia Spotlight [28]. The last one also enables interconnection of the concepts found in the content to DBpedia entities, thus, permitting to enrich even further the text with Open Data. The NLP algorithms used are the ones described in [29]. The resulting tags are filtered (duplicates and stopwords are removed) and then proposed to the users in order to aid them enrich the content with the ones that better represent the inserted text.
2. **Content recommendation based on content similarity:** Two different resources that contain similar words, most probably refer to similar concepts. This is the basis of the content similarity based recommendation systems, which is implemented through the robust and efficient *tf.idf* approach. Therefore, whenever a user is reading a resource, he also finds available the most similar documents.
3. **Content recommendation based on tag similarity:** Given that tags describe the main concept of a resource in a very synthetic way, the same algorithm as in the previous case (*tf.idf*) is employed to recommend to the users similar content, but this time the comparison is not done between the whole text content but only between the tag sets.
4. **Semantic based recommendation:** The semantic relations that the CLIIP framework discovers are used in this recommendation system in order to propose similar resources based on the DBpedia's semantic categorisation. In particular, the semantic similarities between terms are exploited in order to show resources based on the tags they contain. This module is based on the semantic search so the example is the same, but this time the three specific folksonomies are searched independently.
5. **Browsing history recommendation:** This recommender system is based on the assumption that users who search for similar resources are interested in similar content. For this reason, whenever a user navigates to a resource, this action is registered and compared to the habits of other users. Common patterns in the browsing path are identified and the resources found in these paths are proposed to the users. For example, if a user visits a resource titled "Ammonium Nitrate" and after that visits a resource titled "Fertilizers", the browsing history recommender will propose the resource "Fertilizers" to users that are currently viewing the "Ammonium Nitrate" resource.

VIII. Conclusions

This work proposed a framework that integrates several technologies for the discovery, analysis, retrieval, and recommendation of Web resources containing HME information. It is envisaged that within such a framework the tools developed not only provide law enforcement and security agencies with the technology they require to tackle the threat from HMEs, but also assist improved collaborative exchanges and sharing of information

[21] http://opennlp.apache.org/index.html

[22] https://developer.yahoo.com/contentanalysis/

[23] http://new.opencalais.com/

[24] http://developer.zemanta.com/

between agencies based within different jurisdictions. It also supports the individual agencies to minimise resources required to undertake the research into information hosted on the Web in relation to HMEs, ultimately leading to a safer society for the citizen. Of course, this is achieved by also taking measures to be balanced against protecting the right to privacy and right to freedom of speech for the citizen, including understanding the constitution rights of citizens of varying international jurisdictions, whilst realising that the information sharing on the Web is indeed borderless. The development of this framework is currently ongoing and several further components will be integrated in the future, including a social media analysis component that will analyse the social media content in order to identify HME-related posts on social networking sites and detect user communities interested in the domain. Moreover, the framework will be evaluated extensively in terms of its usability, effectiveness, and efficiency by law enforcement and security agencies personnel, as well as by domain experts, in large-scale user studies that will take place in the context of the activities of the HOMER project.

ACKNOWLEDGMENT

This work was supported by the HOMER (312388) FP7 project partially funded by the European Commission.

REFERENCES

[1] A. Stenersen, "The Internet: A Virtual Training Camp?" Terrorism and Political Violence, vol. 20, 2008, pp. 215–233

[2] A. D. de Groot, "Methodology. Foundations of inference and research in the behavioural sciences," 1969, Mouton, The Hague, The Netherlands.

[3] R. A. Baeza-Yates and B. A. Ribeiro-Neto, "Modern Information Retrieval - the concepts and technology behind search", Second edition, 2011, Pearson Education Ltd., Harlow, England.

[4] R. Dingledine, N. Mathewson, and P. Syverson, P. "Tor: The second-generation onion router", In Proc. of the USENIX Security Symposium, 2004, pp. 303–320.

[5] H. Chen, " Dark Web: Exploring and Data Mining the Dark Side of the Web", 2011, Springer.

[6] C. Aliprandi, A.E. De Luca, G. Di Pietro, M. Raffaelli, D. Gazze, M.N. La Polla, A. Marchetti, and M. Tesconi. "CAPER: Crawling and analysing Facebook for intelligence purposes", Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2014, pp. 665-669.

[7] M. Lupu, M. Salampasis, and A. Hanbury, "Domain Specific Search," in Professional Search in the Modern World, 2014, pp. 96-117.

[8] S. Oyama, T. Kokubo, and T. Ishida, "Domain-Specific Web Search with Keyword Spices," IEEE Transactions on Knowledge and Data Engineering, vol. 16 (1), 2004, pp. 17 – 27.

[9] C. Kohlschütter, P. Fankhauser, and W. Nejdl, "Boilerplate Detection Using Shallow Text Features," Proc. Third ACM International Conference on Web Search and Data Mining (WSDM 2010), 2010, pp. 441-450.

[10] B. Stein, M. Koppel, and E. Stamatatos, "Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection," ACM SIGIR Forum, vol. 41(2), 2007, pp. 68-71.

[11] D. Fetterly, M. Manasse, and M. Najork, "On the Evolution of Clusters of Near-Duplicate Web Pages," Proc. 1st Conference on Latin American Web Congress (LA-Web 2003), 2003.

[12] C. Olston, and M. Najork, "Web Crawling," Foundations and Trends in Information Retrieval, vol. 4(3), 2010, pp. 175-246.

[13] T. Tsikrika, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris, "Focussed Crawling of Environmental Web Resources: A Pilot Study on the Combination of Multimedia Evidence," Proc. 1st International Workshop on Environmental Multimedia Retrieval (EMR 2014), in conjunction with the ACM Conference on Multimedia Retrieval (ICMR 2014), 2014, pp. 61-68.

[14] G. Pant, and P. Srinivasan, "Learning to Crawl: Comparing Classification Schemes," ACM Transactions on Information Systems, vol. 23(4), 2005, 430-462.

[15] J. Shakes, M. Langheinrich, and O. Etzioni, "Dynamic Reference Sifting: a Case Study in the Homepage Domain," Proc. 6th International World Wide Web Conference (WWW6), 1997, pp. 189-200.

[16] G. Agarwal, G. Kabra, and K.C.C. Chang, "Towards Rich Query Interpretation: Walking Back and Forth for Mining Query Templates," Proc. 19th ACM International Conference on World Wide Web (WWW 2010), 2010, pp. 1-10

[17] J. R. Quinlan, C4.5: Programs for Machine Learning. Elsevier, 1994.

[18] M. Porter, "An algorithm for suffix stripping", Program, vol. 14(3), 1980, pp. 130−137.

[19] E. Apostolidis and V. Mezaris, "Fast Shot Segmentation Combining Global and Local Visual Descriptors", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[20] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf", 2011 IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2564–2571.

[21] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, 2004, pp. 91–110.

[22] K. Van de Sande, T. Gevers, and C. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 32(9), 2010, pp. 1582–1596.

[23] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," Computer Vision & Image Understanding, vol. 110(3), 2008, pp. 346–359.

[24] F. Markatopoulou, N. Pittaras, O. Papadopoulou, V. Mezaris, and I. Patras, "A Study on the Use of a Binary Local Descriptor and Color Extensions of Local Descriptors for Video Concept Detection," Proc. 21st International Conference on Multimedia Modeling (MMM 15), 2015.

[25] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating Local Image Descriptors into Compact Codes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34(9), 2012, pp. 1704-1716.

[26] Y.K. Lee, S. Yoo, K. Yoon, and B. Berra, "Index structures for structured documents," Proc. first ACM International Conference on Digital Libraries, ACM, 1996, pp. 91-99.

[27] K. Bratanis, D. Bibikas, and I. Paraskakis, "Enabling Cross-Language Intelligent Information Processing in Multilingual Social Networks", Technical Report, Thessaloniki, Greece: South East European Research Centre (SEERC), 2012.

[28] P. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia Spotlight: Shedding Light on the Web of Documents," Proc. 7th International Conference on Semantic Systems, 2011, pp. 1-8.

[29] E. Kargioti, D. Kourtesis, D. Bibikas, I. Paraskakis, and U. Boes, "MORMED: Towards a Multilingual Social Networking Platform Facilitating Medicine 2.0," Proc. XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010, Springer Berlin Heidelberg, 2010, pp. 971-974.