IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 25, 2017, accepted October 25, 2017, date of publication November 2, 2017, date of current version December 5, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2768564

A Fuzzy Ontology and SVM–Based Web Content Classification System

FARMAN ALI[1], PERVEZ KHAN[2], KASHIF RIAZ[3], DAEHAN KWAK[4], TAMER ABUHMED[5], DAEYOUNG PARK[1], KYUNG SUP KWAK[1], (Member, IEEE)

[1]Department of Information and Communication Engineering, Inha University, Incheon 22212, South Korea
[2]Department of Electronics Engineering, Incheon National University, Incheon 406-772, South Korea
[3]University of the Punjab, Gujranwala 52250, Pakistan
[4]Department of Computer Science, Kean University, Union, NJ 07083 USA
[5]Department of Computer Engineering, Inha University, Incheon 22212, South Korea

Corresponding author: Kyung Sup Kwak (e-mail: kskwak@inha.ac.kr)

This work was supported by a National Research Foundation of Korea funded by the Korean Government through the Ministry of Science, ICT, and Future Planning under Grant NRF-2017R1A2B2012337.

ABSTRACT The volume of adult content on the world wide web is increasing rapidly. This makes an automatic detection of adult content a more challenging task, when eliminating access to ill-suited websites. Most pornographic webpage–filtering systems are based on n-gram, naïve Bayes, K-nearest neighbor, and keyword-matching mechanisms, which do not provide perfect extraction of useful data from unstructured web content. These systems have no reasoning capability to intelligently filter web content to classify medical webpages from adult content webpages. In addition, it is easy for children to access pornographic webpages due to the freely available adult content on the Internet. It creates a problem for parents wishing to protect their children from such unsuitable content. To solve these problems, this paper presents a support vector machine (SVM) and fuzzy ontology–based semantic knowledge system to systematically filter web content and to identify and block access to pornography. The proposed system classifies URLs into adult URLs and medical URLs by using a blacklist of censored webpages to provide accuracy and speed. The proposed fuzzy ontology then extracts web content to find website type (adult content, normal, and medical) and block pornographic content. In order to examine the efficiency of the proposed system, fuzzy ontology, and intelligent tools are developed using Protégé 5.1 and Java, respectively. Experimental analysis shows that the performance of the proposed system is efficient for automatically detecting and blocking adult content.

INDEX TERMS Data mining, semantic knowledge, fuzzy ontology, SVM, adult content identification.

I. INTRODUCTION

Currently, the amount of adult (pornographic) content on the Internet is increasing rapidly. It is evident that the existing adult content filtering systems cannot efficiently classify the context of webpages to block ill-suited content. A huge number of adult webpages on the Internet are freely available to all users, which can damage the mental and physical health of teenagers [1]. It also creates problems for parents wishing to keep children away from these webpages [2]. In addition, some webpages contain a huge amount of combined data related to healthcare (information on diseases, mental health, and physical fitness) and sexual knowledge (medicine for sexual health, birth control, treatment during pregnancy, etc.). Keyword-based searching is a renowned form of search engine query to easily retrieve the data of these webpages [3]. It retrieves the data by comparing the query keywords with web content words. However, the existing adult content filtering systems are inefficient at detecting whether the webpage is about pornography or medicine.

At present, the detection of adult content is based on uniform resource locator (URL) filtering, image filtering, and dynamic filtering mechanisms. URL filtering methods use URL blacklists and do not evaluate the content of a webpage, which increases the possibility of a wrong decision. Image filtering techniques might identify medical-related images as adult images, and they have a low performance capability. A dynamic filtering system analyzes the content of a webpage using various algorithms [4]. However, the extraction of meaningful keywords to classify the content of a webpage is a challenging task for these systems. The identification of webpage type without access to internal keywords is a research question in itself, as well as necessary for other

analyses. The existing systems employ common features to block access to material they determine to be immoral. In addition, the present systems acknowledge only broad categories of adult content (*sedition*, *obscenity*, etc.). Nevertheless, meaningful keywords are kept secret and can be altered without notice, and may differ from common features. Most of the existing systems use naïve Bayes, the bag-of-words model, or a classical ontology to retrieve and classify objectionable material [5]–[7]. These systems are unable to find valuable content and remove irrelevant content. Furthermore, Platform for Internet Content Selection (PICS) has been used to classify webpages into a whitelist and a blacklist [2], and uses a general format for data labels and vocabulary. However, the PICS-based approach is inadequate at effectively classifying medical webpages because of its semantic limitations. Two main strategies are used to address this issue. The first strategy is to categorize the URL to provide accuracy and speed. The second strategy is classification of data labels, which are descriptions of the resources. In addition, most information retrieval and extraction is based on a classical ontology [5], [7], [8]. A classical ontology–based system is insufficient, and can extract useful data from the Internet only to a limited extent. Therefore, an ontology with fuzzy logic is considered an efficient technology for adult content filtering systems in order to retrieve meaningful data from the merged web content.

To solve these problems, this paper proposes a support vector machine (SVM) and fuzzy ontology–based adult content detection system. The proposed fuzzy ontology provides semantic knowledge for ill-suited content detection, and the SVM removes irrelevant content. In this paper, we analyze the context of webpages without image filtering to prevent a low performance capability in image scanning. The overall process of the proposed system is divided into three parts: adult content (features) extraction, fuzzy ontology–based knowledge representation, and identification of webpage type. The main contributions of this research are the following.

- In the adult content–filtering phase, the proposed system adds URLs to a blacklist or a whitelist by making comparisons with an ontology URL list and a URL blacklist available online. Speed and accuracy are key advantages of this phase. Then, the web contents are intelligently analyzed to detect adult webpages by extracting informal data from the webpage content. If the amount of informal data exceeds a predefined threshold value, then access to the webpage will be restricted systematically. Sometimes, this technique over-blocks webpages due to a lack of semantic understanding. For example, a system may block a webpage as having unsuitable data because it contains medical terms. However, the proposed system overcomes this problem because it describes the concepts of adult content using a fuzzy ontology. It identifies pornographic webpages accurately and reduces misunderstanding from misrecognizing medical content as adult content.
- An unsupervised linear technique is employed to extract meaningful keywords from adult webpages and to filter out irrelevant words.
- The fuzzy ontology contains all visited URLs, supporting words, medical terms, pornographic terms, and normal words. It provides a semantic web rule language (SWRL) rule–based knowledge platform for specific data extraction and intelligent classification to automatically identify the webpage type (adult website, medical website, or normal website).
- The proposed fuzzy inference layer and semantic knowledge are employed to detect and block adult content.

The rest of this paper is structured as follows. Section 2 summarizes the existing research. The basic concepts of a fuzzy ontology are defined in Section 3, whereas Section 4 briefly explains the overall scenario and internal process of the proposed system. In Section 5, the experimental work and results are discussed. Finally, Section 6 concludes this paper.

II. RELATED WORK

Information extraction and pornographic content filtering are hot topics in the field of information engineering research [1], [4], [9]–[14]. The increase in adult websites on the Internet has made web filtering a more challenging task. Most pornographic website filtering systems are unable to filter data efficiently to prevent teenagers from accessing them. As a solution, comprehensive technological work is required to extract and filter web data intelligently and deny access to ill-suited webpages systematically. One possible method of webpage filtering is to record the URLs of ill-suited websites. The main advantage is speed. However, a URL-based filtering system does not work perfectly every time, since many URLs do not present the actual information. To handle this limitation, web content-based filtering and blocking techniques are required to filter webpages competently. Different links and Hypertext Markup Language (HTML) tags in webpages contain a lot of information that can be used for filtering. Similarly, images and video-based filtering and blocking techniques are also used to stop access to unsuitable webpages. These techniques use various neural network algorithms to detect and block objectionable images in webpages [15]–[17].

A machine learning algorithm–based system was presented to classify webpage URLs [16]. This system uses the Open Directory Project (ODP), which contains all blacklisted URLs. During execution, every URL is compared with the ODP blacklist to filter webpages. Two techniques, called token and n-gram, are used in the system for URL classification into pornographic features or non-pornographic features. However, there are limitations. Sometimes, a URL phrase does not reflect the webpage's real data. Immoral data are hidden in the content of the webpages. A URL-based filtering system is unable to block inappropriate webpages. A naïve Bayes technique was used for supervised learning and classification of weblogs by splitting the URL into tokens [15]. This system solved different problems in

the spam blog (splog) system. Most of the URL phrases, like http://adult-videompegs.blogspot.com, contain punctuation segmentations and combined words, which cannot be filtered by the splog system. To overcome the URL phrase problem, a system was developed to categorize weblogs as spam or good. The system compares the URL with a URL repository. If the URL is not a part of the repository, then the URL is added, and the weblog category is updated. An unsupervised statistical technique was proposed to extract features from URL strings [17]. These features are employed to build URL patterns. The URL is tokenized, and feature values of the tokens are then calculated from the training set and subset to build the pattern. For example, a URL like http://www.abc.com/pro/test = 24 gives the pattern http://www.abc.com/∗/dp- bc = ∗, which represents all links to the specific location. Two mechanisms were presented to blacklist objectionable web content data in yet another system [18]. The first mechanism is based on content analysis, and the second is behavioral attributes of the queried addresses. The content analysis–based mechanism focuses on a set of IP addresses, and the behavioral analysis–based system focuses on client requests for listed information on the server. The optimization of blacklisting is achieved, and spam is avoided using these two methods. A technique called the multi-level counting Bloom filter was presented to describe network-based URL filtering (NUF). This system restricts pornographic traffic over a network [19]. NUF is used between client and server for URL filtering. Whenever a request for URL access is generated from the client side, NUF is activated to evaluate the URL. If the URL is in the to-be-blocked cache, the system sends a block message to the web server. Otherwise, the URL will be registered for future reference by the filtering server. An increase in the cache reduces the traffic rate and bandwidth cost. However, this technique cannot efficiently block pornographic webpages on the Internet because there are millions of obscene URLs to analyze. A search-intent method based on a pornographic material blacklist is used for joint cyber porn filtering [20]. This framework finds new uploaded pornographic webpages. This system first checks URLs with any categorical keywords that might be suspect, and the URL is then added to a blacklist if the suspicion proves correct. The system achieved a high blocking rate with respect to time and maintenance. A named entity recognition (NER)-based system was presented to filter web content [21]. This system uses simple techniques like keyword matching or URL blocking that limit their effectiveness at filtering immoral content. A model lexical NER system helps in demonstrating the web content filtration as it takes training pages for the web. The NER system extracts and tokenizes words to assign a weight vector using a support vector machine (SVM). It improved the classification of webpage text. Important issues were discussed regarding access to unsuitable websites at home, in school, and in organizations [22]. It becomes increasingly difficult to prevent employees and children from accessing these pages.

During the past few years, content-based filtering has been used for webpage filtering [23], [24]. SVM with K-nearest neighbor is employed to remove noisy data in the training set [25]. This system achieved accuracy of 87% with a training set of 1400 webpages. A naïve Bayes approach was used to classify webpages [26]. The system extracts features from HTML tags, and the central limit theorem is then employed to find the weight of the features. This process uses different functions; a naïve Bayes theorem calculates the probabilities of various data sets and events, and a Gaussian model classifies the web documents. These functions use the plain text in the HTML document to find frequencies and weight. A dynamic threshold and speculative aggregation technique is used for host blacklisting and thresholding [27]. This system describes the blacklisting policy with the help of local information, including usage patterns, network usage visibility, and global usage. A threshold is fixed; if spam crosses that threshold, the email server will be added to the blacklist. The dynamic threshold approach computes the ratio of emails with a spam trap to identify the server for blacklisting. Border Gateway Protocol (BGP) is used to swap routing information between autonomous systems and identify logical and organizational boundaries. A higher-level recommender insert strategy was presented to suggest a set of indexes for an underlying external universal recommender [28]. It matches keywords with an internal dataset of the organization, which consists of some specific keywords. This recommender system has the ability to monitor activities in software development. It checks the webpage content and finds the index title, uniqueness of the URL, and material relevancy. It overcomes cross-organizational privacy issues. A web content classification system was presented to update a blacklist with inappropriate websites [20]. This approach performs a classification and incremental update process. The web swarming sub-process collects pornographic material using pornographic keywords in search engines. These materials are then classified into three categories by assigning values to each page. This system removes HTML tags, finds the encoding language, selects keywords from the content data, and constructs a feature set. A specific value is calculated and a category level is assigned to these webpages. The accuracy rate of this system is 97.11%. An intelligent web filtering system called XFighter was presented to filter web content [29]. XFighter consists of three main components; an access control database, an offline classification agent, and an online filtering agent. The offline classification agent discovers objectionable webpages. The access control database contains information regarding URLs. The online filtering agent monitors online browsing and performs blocking. A descriptive model based on Bayesian classification was presented to block pornographic webpages [30]. This model is based on structured modeling (images, links), textual modeling (HTML tags, metadata), and term variation. Bayesian classification delivered 99.1% accuracy on both pornographic and non-pornographic content. Skin detection is usually used as the most common parameter for detecting and blocking

obscene images. Adult image classification methods use two kinds of filter; an adult image filter and a harmful symbol filter [31]. The adult image filter uses a statistical model for skin detection and a neural network for adult image classification. The experimental results with both filters showed promising performance. A novel framework for webpage splitting handles three categories of webpages. These is a continuous text classifier, a discrete text classifier, and a fusion-of-images classifier [32]. These classifiers provide a decision symbol (porno or non-porno) to the web browser.

Nowadays, the ontology is extensively employed in the field of web filtering and information extraction. An ontology shares domain information between individuals. A classical ontology for information content filtering was presented to clarify the blurred definitions of the common concepts in the specific domain and the relationships between concepts [33]. This ontology defines the content information in the form of keywords using concepts and their relationships. It also defines the content filter inference rules with constraint sets for analyzing the context of the word of concern (WoC). An ontology-based integrated approach was presented for information retrieval and filtering [8]. This system provides a tool for acquiring webpages that hold relevant information about the domain concepts. It uses two techniques for information retrieval and acquisition. The first technique allows the use of information contained in the ontology. The second technique is automatic and a domain-independent ontology-learning method for machines. An ontology-based filtering mechanism was proposed to solve two problems of web usage mining related to the pattern analysis phase, such as pattern retrieval and pattern interpretation [34]. The pattern retrieval problem manages a large set of patterns. The difficulty with these pattern interpretations is that they analyze the relevance of patterns with regard to the domain. A multipurpose ontology-based system was presented to enhance the representation of relevant information about contextual conditions and the retrieval process [35]. This system defines the development of advanced features and the enhancement of personalization. Moreover, a literature summary and comparison of different techniques is shown in Table 1.

A discussion of previous studies was observed in depth, and it was determined that some of the research provides very interesting background, and some research is flawed in the area of information extraction and adult content filtering. Most of the systems cannot achieve the correct results from intensively blurred data. In addition, the existing systems extract the data from webpage content to a limited extent, and are unable to detect and block pornographic content perfectly. The proposed fuzzy ontology–based knowledge system is a novel attempt to develop automatic web content filtering. In the proposed system, an ontology is used to provide high-level knowledge representation for webpage filtering to make the decision-making technique more intelligent. The proposed ontology represents four types of information: features and keywords related to adult content, medically related words, normal words, and supporting words. The user's request for a URL is assigned to the web crawler, the web content data are retrieved, and keywords and values are then extracted from the retrieved data using the fuzzy ontology to calculate the indicator value for the decision-making system. Furthermore, a knowledge- and SWRL-rule–based ontology defines rules (if-then) for the decision system. The fuzzy inference layer is merged with the ontology to easily extract information and identify web type.

III. FUZZY ONTOLOGY

In this section, the concept and terminology of an ontology is defined before moving on to fuzzy ontologies. An ontology is used to share knowledge of a specific domain, which can be understood by both machine and human [44]. It shares specific domain information for reuse, instead of remodeling the information [45]. The ontology is becoming the cornerstone of the semantic web. It retains information and contributes semantic interoperability between applications [46]. There are four main elements to an ontology: classes, concepts, instances, and relationships. There are three kinds of Web Ontology Language (OWL): OWL-Lite, OWL-Descriptive Logic (DL), and OWL-Full. OWL-Lite classifies a hierarchy with simple constraints. It supports cardinality constraints using the values 0 or 1. OWL-Lite is easier to use than OWL-DL. However, OWL-DL constructs all OWL languages under certain restrictions. OWL-DL is famous in the field of research due to its correspondence with description logic. OWL-Full supports Resource Description Framework (RDF) with no computational guarantees. However, OWL and RDF were used to develop the proposed ontology [47]. Mathematically, an ontology can be defined as follows [46]:

$$\widetilde{\text{Ontology}} = (\text{C}, \text{P}, \text{R}, \text{V}, \text{V}_\text{c}) \tag{1}$$

where C, P, R, V, and V_c stand for concepts, properties of concepts, relationship among concepts, value of concepts, and the constraint value of properties, respectively. The proposed ontology represents information about a specific domain (a URL with web content data). However, there is no general way to make an ontology, and researchers develop ontologies based on their own needs. There are two types of ontology: the classical ontology and the fuzzy ontology. In a classical ontology, the value of a concept is crisp, but most of the actual system takes fuzzy terms [48]–[50]. As a result, a crisp ontology with fuzzy logic is an effective way to keep the system from imprecise data and arrive at the indicator value. Zadeh launched fuzzy set theory in 1965 [51]. The essential features of typical set theory are that something either belongs to a proper set or does not. There is no ''membership value'' notion in a typical set. However, fuzzy set components with a proper set exist between 0 and 1, which is called the degree of membership [52]. Fuzzy set theory simplifies the concepts in crisp set theory to present imprecise boundaries, such as a normal, medical, or pornographic webpage. Mathematically, membership function μ_F is used to define a fuzzy set, F, over the universe of discourse, A, which declares element A in the

TABLE 1. Literature summary and comparison of different techniques.

Systems	Purposes	Techniques	Dataset	Accuracy
[16]	URL-based webpage classification	N-gram	4,167	94%
[19]	Cache mechanism-based blocking of undesirable URLs	Multi-level counting Bloom filters	Nil	91%
[36]	URL-based detecting of malicious websites	Naïve Bayes and SVMs	15000	90%
[37]	URL-based pornographic web content filtering	K-nearest neighbor	Nil	96%
[25]	Filtering based on hypertext classification	K-nearest neighbor and SVM	1,400	87%
[28]	Blocking webpages by using URLs and content	Web search-centric approach	Nil	84%
[38]	Prevent webpage malicious code execution and filter contents	Prophiler	18,939,908	85%
[39]	Objectionable content blocking	Early decision algorithm	6,000	89%
[40]	Classifying webpages by using their contents	Early decision algorithm	1250	93%
[41]	Classifying videos by using visual features	Bag-of-visual-features (BoVF)	Nil	93.%
[42]	Image-based detection of normal and obscene videos	Weighted support vector machine (WSVM) classifier	1,100	94.%
[22]	Identifying undesirable webpages on a client PC	Cerebellar model articulation controller (CMAC)	400,000	85.%
[43]	Content-based blocking of malicious websites	Multiple string matching based (Wu-Manber-like)	Nil	80%
[9]	Making social network free of harmful content	Radial basis function network model (RBNF)	Nil	.....

interval [0, 1]:

$$\mu_{\mathrm{F}}(\mathrm{A}) : \mathrm{A}' \rightarrow [0, 1] \tag{2}$$

In the above expression, μ_{F} presents the membership degree to which A' belongs to A. A' is a complete member of set A if $\mu_{\mathrm{F}}(\mathrm{A}) = 1$, and a partial member of set A if $\mu_{\mathrm{F}}(\mathrm{A})$ is in the interval [0, 1] (e.g. 0.54). The fundamental basis of the proposed adult content detection system is a fuzzy ontology, which not only offers semantic knowledge to compute the indicator value, but also uses queries to retrieve the needed information from the ontology.

IV. DEVELOPMENT OF FUZZY ONTOLOGY AND SVM-BASED ADULT CONTENT FILTERING

The main aim of this research is to present a computational method for automatically detecting and blocking adult content. The detection of ill-suited webpages is based on web content filtering of the pornographic webpages identified accurately, and reducing misunderstanding from recognizing a medical webpage as an adult content webpage. In this section, the internal process and architecture of the proposed system is presented, as shown in Fig. 1. For simplicity, the architecture is divided into three steps, as follows:

- Adult content (feature) extraction
- Fuzzy ontology–based knowledge for web categorization
- Fuzzy inference layer-based indicator value computation for adult content blocking

This study includes three different types of lists with 341,362 URLs. Some of the URLs contain more adult content than anything else; we consider these URLs as a blacklist. In addition, URLs that comprise sensitive topics and more likely to

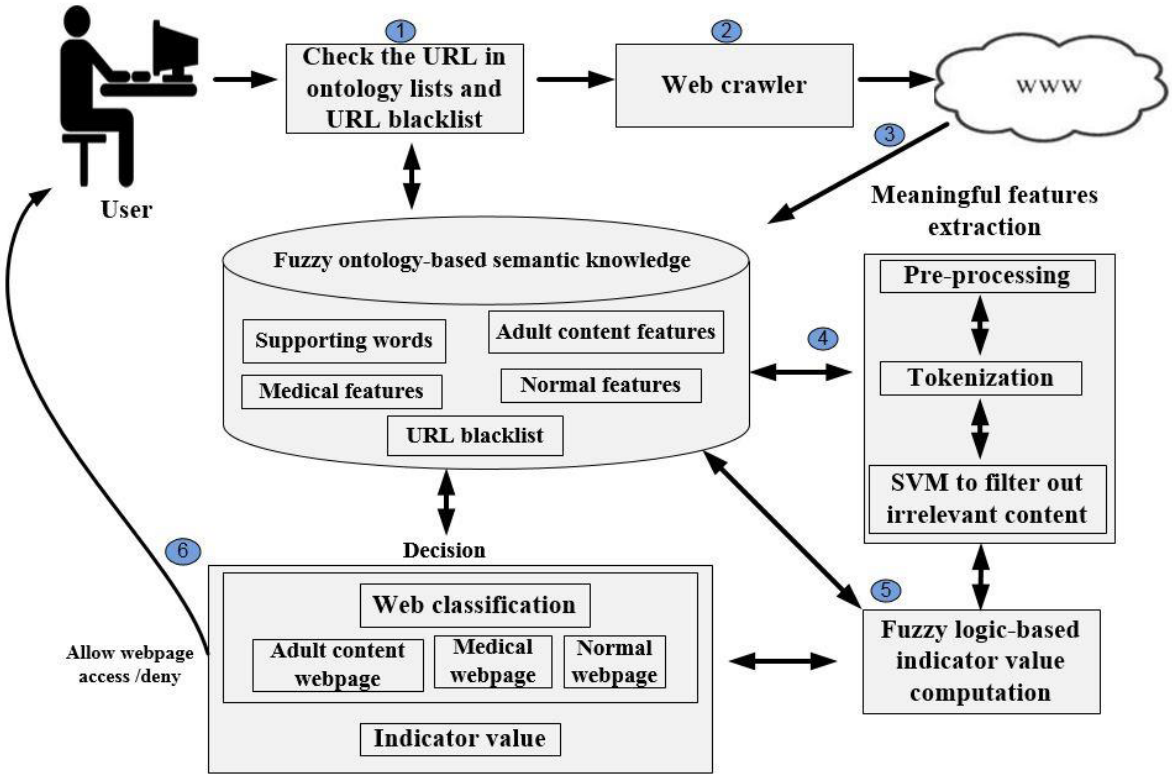

FIGURE 1. The proposed system architecture.

be pornographic than average are also considered a blacklist. Generally, users enter a URL to retrieve specific information. The proposed system compares the URL with fuzzy ontology information and the URL blacklists to make a clear distinction between normal and pornographic URLs (task 1 in Fig. 1). If the URL already exists on a URL blacklist and in ontology classes, then the system does not need to retrieve the web content to identify the web category. However, it directly sends the category of that URL from the semantic information to the decision part of the proposed system. Speed and accuracy are key points of this step. After confirmation of the URL in the ontology lists and blacklists, the proposed web content filtering system sends the URL to the web crawler (task 2 in Fig. 1). The proposed system extracts features (pornographic terms) along with meaningful keywords from the web crawler and sends it to the semantic knowledge–based system along with the URL for further processing. The proposed semantic knowledge–based system handles any kind of practical scenario associated with pornographic content filtering and blocking. It stores visited URLs, supporting words (good words and bad words), pornographic words, medical words, and normal words, which can be reused when the system retrieves a similar webpage or a new webpage. This information was collected from the Internet and categorized manually. The gathering of information from the Internet was a major task and helps to accelerate the design process of the semantic knowledge–based system for adult content filtering. The proposed system contains analyzed information: for example, normal-string (data property) is a good-word of a normal-words list (superclass); evaluation-value (data property) is an indicator-value for the web type, adult content, and normal words (superclasses); and primary-source (data property) is a first-priority string of normal words (superclass). If the current URL is not on the list of ontology classes and the URL blacklists, the system then retrieves the web content from the web crawler and delivers it to the web content extraction unit (task 3 in Fig. 1). The web content extraction unit extracts useful keywords from the webpage content to find the webpage type (task 4 in Fig. 1). The web content is pre-processed to eliminate stop-words, and tokenize the web content. After pre-processing, the SVM is applied to identify useful keywords and filter out unrelated keywords. It classifies the webpage content by using a linearly separable hyperplane with binary categorization [53], [54].

The proposed system gets a value for each extracted keyword from the ontology to evaluate the webpage. Fuzzy logic along with ontology knowledge then uses these keywords to classify the website into one of three categories: adult, medical, or normal. In addition, it also assigns these words, along with values, to the fuzzy inference layer unit. This unit uses fuzzy logic to compute the final indicator value regarding the current webpage (task 5 in Fig. 1). The system forwards the indicator value to the decision unit. The decision unit automatically responds to the system to allow or block

access to the webpage (task 6 in Fig. 1). Protégé OWL was used to develop the proposed ontology [55]. DL and Simple Protocol and RDF Query Language (SPARQL) queries are employed to retrieve the web content from semantic knowledge for the indicator value computation [56], [57]. These queries retrieve the instance of each class, which shows the internal relationship among classes. First, a classical ontology is developed and a fuzzy OWL plugin is employed to convert the simple ontology into a fuzzy ontology. The fuzzy ontology is a semantic knowledge representation that generates a web content filtering lexicon. It expedites the proposed adult content filtering system to help categorize the web content and compute the correct indication value for the webpage. The internal processes of the proposed system are discussed step-by-step in the next sections.

A. ADULT CONTENT (FEATURE) EXTRACTION

The World Wide Web contains an enormous amount of information. People can easily access this information with just a simple text query. However, access to adult content pollutes teenagers' minds and risks sexual abuse. All webpage content needs to be filtered, specifically for detecting and blocking pornographic webpages. One possible way of webpage classification is assessment of the URL. However, URL-based classification will not work every time, since many URLs do not reflect the actual webpage content. To handle this limitation, the proposed system uses web content–based filtering and blocking techniques. A support vector machine is employed to filter out irrelevant content, and fuzzy logic is then used to classify the webpage into three categories: adult webpage, medical webpage, and normal webpage [54].

To achieve an appropriate and exact ontology lexicon, the adult content (feature) extraction unit is divided into two steps: URL-based analysis, and SVM-based content analysis and feature extraction. These steps are performed in series. Eclipse with the Protégé OWL ontology was used to make the proposed adult content filtering system. Eclipse provides essential functionality to run or create additional modules. These modules are plugins that represent the smallest units of an Eclipse function. The second version of the Jena application programming interface (API) is used to retrieve meaningful information from the fuzzy ontology. The Jena API provides reasoners in order to support inferences. These reasoners are transitive, Resource Description Framework scheme (RDFS) rule, OWL, OWL Mini, OWL Micro, and a generic rule.

1) URL-BASED ANALYSIS

The most common protocols are http, https, and ftp, which are used by people to locate a website or server. A URL is a human-readable text string. The URL string *http://www.breastcancer.org/* is employed to understand the URL-based filtering process precisely. After assigning the user request to the web crawler for URL access, the proposed system automatically removes its prefix and postfix, such as http://, .com, .edu, etc. The string is then compared with the blacklist and the lexicon dictionary of the fuzzy ontology. If the URL contains safe domains (such as .edu or .gov), the system does not compare the URL with the ontology lexicon, but allows the user to access the URL. Beyond that, if the system retrieves any unsuitable word, then the URL string is compared with the lexicon dictionaries of both the blacklist and the medical webpage list. It confirms whether the extracted words are related to the field of medicine or not. It may contain medical words or other languages, e.g., breast cancer. The proposed system processes any ''breast'' words as unsuitable words, but the word cancer is related to medicine. Therefore, if the URL string holds medical words, the system then allows users to access it with a warning message. This reduces the blocking probability of medicine-related websites.

2) SVM-BASED CONTENT ANALYSIS AND FEATURE EXTRACTION

In this phase, the webpage is examined and adult content is extracted by employing pre-processing and feature extraction. In the pre-processing step, the webpage content is converted into HTML format. Webpages contain different types of scripts, such as the style sheet, the title, and metadata information. The proposed system extracts this information, and morphological analysis is then employed to identify the various forms of words by employing a lexicon. A lemmatization procedure with morphological analysis verifies the lemmas of the words in the <title> tag. For example, the tag *<title>Because every tumor is unique | Genetic Testing for Breast Cancer, Breast cancer New York, Targeted Therapy for Breast Cancer, Breast Cancer Los Angeles, Rates of Breastfeeding Recurrence, beautiful breastfeeding for babies, Molecular Subtype</title>* contains the words *testing*, *targeted*, and *rates*. These words have the basic forms *test*, *target*, and *rate* as their lemmas. The system uses lemmas of these words for further processing [58]. After the lemmatization process, the system invokes a supporting WordNet to clarify ambiguities in the <title> tag. It offers an opportunity to choose a suitable meaning for the keywords. The tokenization process splits a compound text into chunks, or tokens. After the tokenization process, the proposed scheme acquires the above <title> tag result in the form of tokens: *because*, *every*, *tumor*, *is*, *unique*, and so on. The results of the tokenization process are stored in the form of an array to compare them with the lexicon and the dictionary of supporting words. After analysis of the title tag, the system fetches descriptions in metadata tags. A word in a description may have different meanings. However, the system uses contextual information to determine the presence of adult content on the webpage. The extracted words may not include negative meanings, but these words are pointed out as useless content when connected with other words. Based on these concerns, we split the keywords into two classes: hidden keywords and obvious keywords. The keywords of the hidden class have no negative meanings by themselves. However, the use of these keywords with other words shows that the text is related to pornographic words with a high probability. Keywords in the

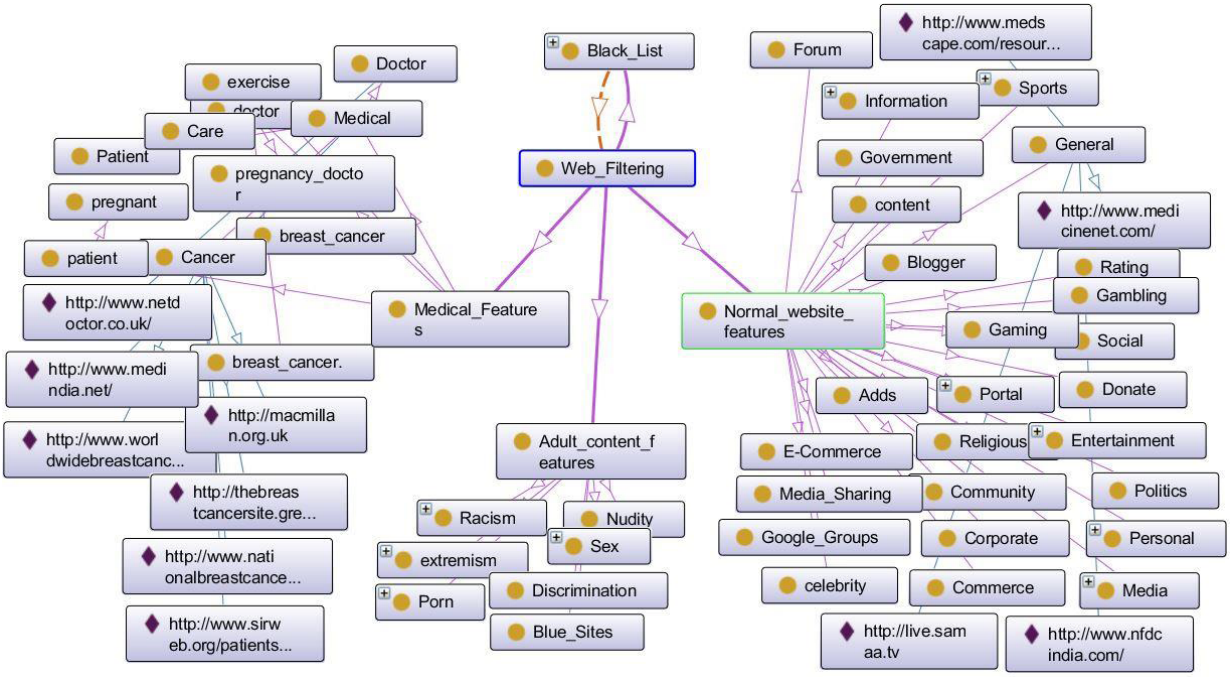

FIGURE 2. Relationships of features in the fuzzy ontology.

obvious class mostly appear in inappropriate text, rather than normal text. After the extraction of these keywords from the webpage content, the SVM with an ontology-based lexicon dictionary is applied to filter out irrelevant keywords.

The second step is feature extraction. After the retrieval of mixed words (normal, medical, and pornographic words), the precision rate is very low. Therefore, the SVM classifier is used to find related features along with meaningful keywords, while unrelated ones are removed. A specific function is applied to identify valuable features. If a content value is greater than 0, that states that the content is related to pornography or medicine; otherwise, the content is filtered out. A simple example is presented to explain the classification of the above tag content using the SVM classifier function: f (webpage content) = 0.1 * Tumor + 0.6 * Breast + 0.5 * Babies – 0.3 * Breastfeeding. The tag content result is f (webpage content) = 0.1 * 1 + 0.6 * 1 + 0.5 * 1 – 0.3 * 1 = 0.9. If f () > 0, then it is positive content (related to pornography and medicine); otherwise, the content is negative [59].

B. FUZZY ONTOLOGY–BASED WEB CONTENT FILTERING

The fuzzy ontology designs the domain knowledge and makes a connection between the information about the web content and the indicator value computation. The proposed fuzzy ontology was developed using seven steps [46]. A domain expert was consulted for every step to obtain accuracy in the experimental results. The relationships among classes (features) in the fuzzy ontology are shown in Fig. 2. The OntoGraf plugin of Protégé was used to develop this graph. The normal website features, adult content features, medical features, and supporting words are described in the fuzzy ontology. For example, in the above-mentioned web content, the words *breast* and*babies,* and the phrase *beautiful breastfeeding* have similarities to the list of medical and pornographic features. Here, *breast* and *beautiful breastfeeding* are declared strings of medical words based on the medical features in the ontology. Furthermore, these words also have negative meanings and can be declared strings with ill-suited words from adult content features. However, the proposed system uses a tokenization process to split the string into small chunks, and a lexicon dictionary is then invoked to replace strings of words with synonyms. For example, *beautiful breastfeeding for babies* is converted to *beautiful feed for babies*. It was already determined that *beautiful* and *feed* are used as normal support words in the proposed ontology. The proposed semantic knowledge uses a semantic score for the strings of words from the web content. The range of semantic scores proposed for normal words is 0.0 to 0.4, for medical words, 0.4 to 0.7 for normal words, and for pornographic words, 0.7 to 1, with a range assigned to the words of the web content during execution. Different kinds of ontology properties and instances (object properties, data properties, property and relationship) are used to assign a semantic score to the words in a string. Object properties interconnect the classes, and these are not linked with a particular class (as they are in object-oriented paradigms) in RDF/OWL. However, these properties are first-class citizens of the ontology and are defined according to our need, which shows the relationships between individuals. Data properties connect individuals with literals. The functional data properties are called *attributes* in the knowledge representation systems. In addition, the proposed ontology uses data properties to show the relationship between individual and literal data. An object property is illustrated as an individual of

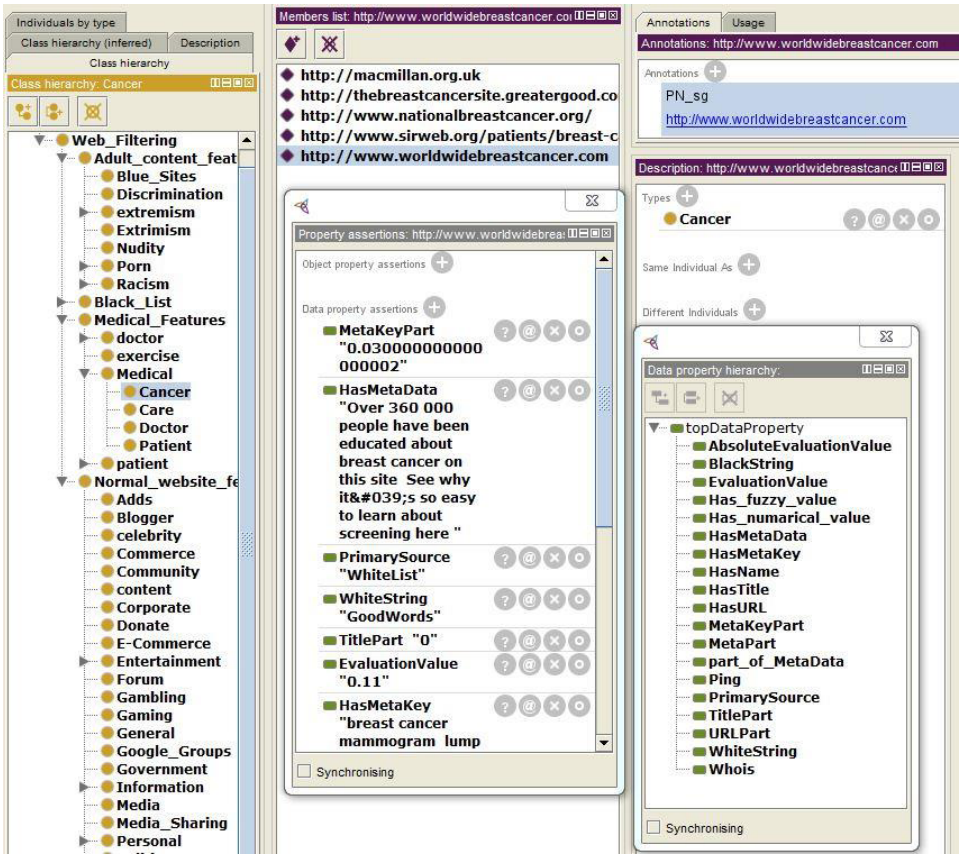

FIGURE 3. Properties and instances of the fuzzy ontology.

the built-in OWL class owl:ObjectProperty, and is a relation between instances of two classes. In the proposed semantic knowledge, object properties have their domain in the class of normal features, medical features, and adult content features, and a range in web content filtering. If one attribute has many domains, then its domain will be the intersection of all the domains. After creating the classes and properties, the fuzzy ontology describes its members. The domain user can easily declare specific properties about individuals. To define an instance (individual), it is important to select the right class and then create its instances. The properties and instances (individuals) of the fuzzy ontology are shown in Fig. 3.

The fuzzy OWL plugin in Protégé 5.1 is a semi-automatic tool that converts a crisp ontology into a fuzzy ontology. This plugin creates fuzzy concepts, fuzzy data types, and fuzzy modifiers. This plugin is applied to represent annotations and linguistic words for the adult content filtering system. These linguistic words are normal words, pornographic words, and medical words, and the range is [0, 1]. The range of these words is systemized using min-max normalization and plotted to the range [0, 1] as described by Ali *et al.* [46]. Fig. 4 shows the declaration of fuzzy data, range mapping, and related information of the class hierarchy. Default reasoner tools, such as FaCT++, the DeLorean reasoner, Pellet, and Hermit, are used to mechanically create the inference results on behalf of the necessary terms in the fuzzy ontology. Moreover, the Protégé tool uses these reasoners for the ontology result, because all fuzzy features are organized as annotations (i.e. fuzzy label annotations). A start tag <fuzzyOWL2> and an end tag enclose the annotation with a fuzzy type attribute to specify the tagging of fuzzy elements. The data type annotations of FuzzyOWL2 for linguistic terms (normal features, pornographic features, and medical features) are the following:

<fuzzyOwl2 fuzzyType="datatype">
<Datatype type="triangular" a="0.0" b="0.275" c="0.55" /></fuzzyOwl2>
<fuzzyOwl2 fuzzyType="datatype">
<Datatype type="triangular" a="0.4" b="0.55" c="0.7" />
</fuzzyOwl2>
<fuzzyOwl2 fuzzyType="datatype">
<Datatype type="triangular" a="0.55" b="0.7" c="1.0" /></fuzzyOwl2>

The proposed system uses a popular reasoner called the DeLorean reasoner. It is employed to retrieve inference results from the ontology [60]. The ontology keeps the web content properties, concepts, and knowledge for each part of the adult content filtering system. The concepts of an adult class can be labeled in fuzzy format as *adult* $\pi\exists$*hisstring*

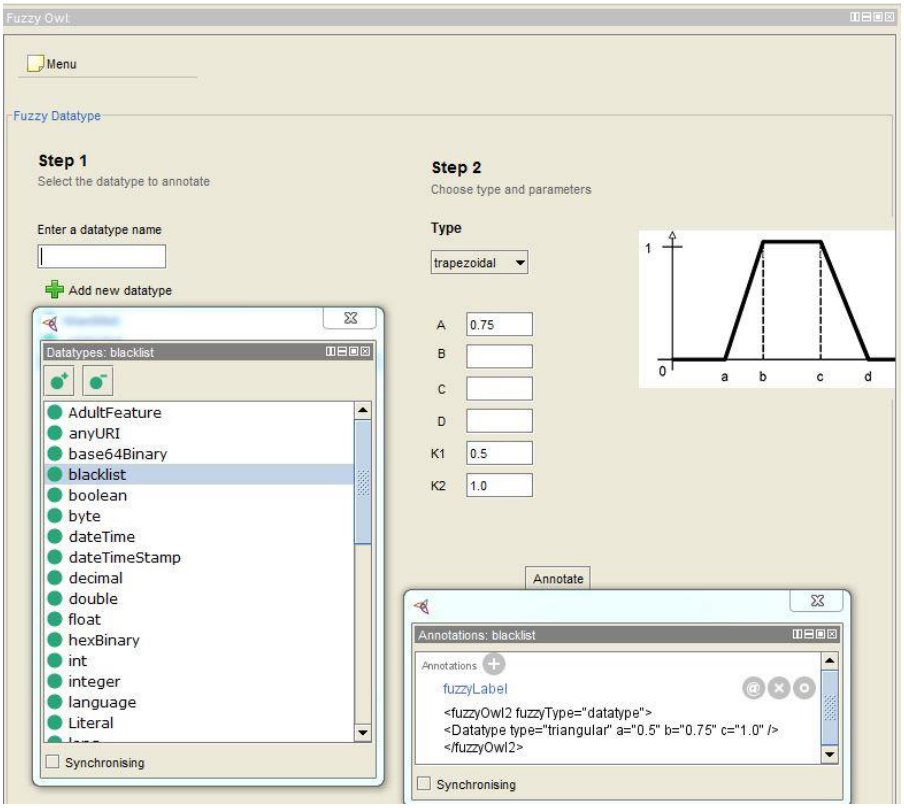

FIGURE 4. Fuzzy data type declaration using the fuzzy OWL plugin.

string_porn. Similarly, the concept of a subclass of adult can be defined as *Sub (adult* $\pi\exists$*hisstring string_porn).* We constructed knowledge and rules using SWRL in the ontology for indicator value calculation. The knowledge side describes the fuzzy terms, fuzzy concepts, fuzzy set membership function, and fuzzy variables, while the rule side defines the fuzzy rules [46]. There are many input string variables (ill-suited words, normal words, medical words, etc.), one output fuzzy variable (adult webpage, normal webpage, or medical webpage), and different fuzzy rules (R1: if the URL string contains ill-suited words, the title tag holds ill-suited words, a description tag comprises ill-suited words, and the meta keywords tag has ill-suited words, then the webpage is pornographic) in the knowledge and rule-based ontology. Each fuzzy variable takes any of three linguistic terms: adult content, normal content, and medical content. The fuzzy decision extracts the information from each webpage and connects its individuals to each semantic term. All individuals are situated in the ontology as webpage_adult, webpage_normal, and webpage_medical. These individuals describe the classification of webpages. The fuzzy OWL plugin of Protégé creates these fuzzy sets and properties. This plugin is unable to competently handle imprecise content. Therefore, during the development of the fuzzy ontology, the crisp ontology is shipped into a simple text editor. The terminology and association of fuzzy logic are manually declared as knowledge representations. Each related concept is categorized into crisp, partially fuzzy, and fully fuzzy.

C. FUZZY INFERENCE LAYER-BASED INDICATOR VALUE COMPUTATION FOR ADULT CONTENT BLOCKING

In the proposed semantic knowledge, a fuzzy knowledge base (FKB) describes a finite set of axioms. FKB = <W, I> comprises a fuzzy box W and a fuzzy box I. Fuzzy box W is the fuzzy ontology classes, FO_c, and fuzzy box I is an instance (individual), FO_I, so FKB = (FO_c, FO_I) [61]. The proposed system uses this information to incorporate the extracted words of the web content and their values to compute the indicator value for a webpage when accessing or blocking it, as shown in Fig. 3. The value of a word depends on the usage condition in the web content. For example, in the lexicon dictionary, the value of *breast* words that describe the feeding style for babies in the description data is 0.5. This value shows that the word is related to medical content. These values are further used during computation of the indicator value. The extraction of words from webpage content and the process of assigning word values are discussed in Section 4.1 and Section 4.2. The fuzzy inference layer integrates the extracted words and their values to verify the web type and block access to adult content. The fuzzy inference layer comprises four elements: fuzzification, inference, defuzzification, and a knowledge-and rule-based ontology. Let us consider the

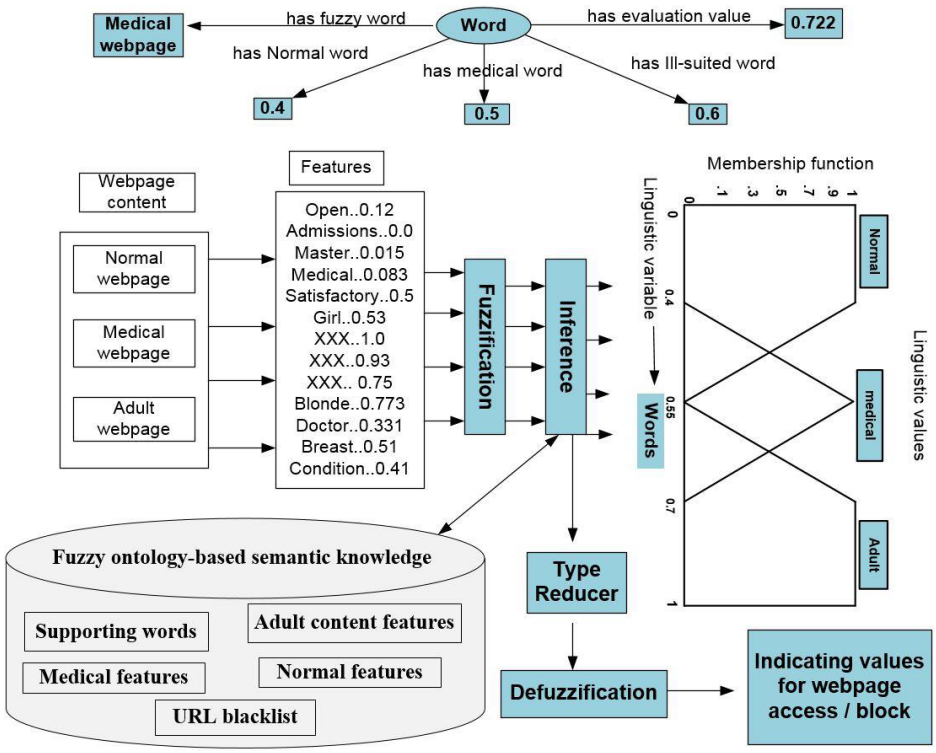

FIGURE 5. Indicator value computation based on the fuzzy inference layer.

above-mentioned content in sections 4.1 and 4.2 related to medical and pornographic features. In the content, the medical words are {medicine, breast, young, and so on}, and for pornography, {breast, babies, XXX, and so on}. First, the fuzzy data are assigned in our proposed technique. The proposed system has various inputs, including the words of the URL string, and the title, body, and meta tags, which are called words of web content. The parameters of web content are classical values that are identified by domain experts. Second, the triangular membership function (MF) defines the membership value for each of the input variables, as shown in Fig. 5. Each variable has three linguistic values: normal, medical, and adult. When the words (medicine = 0.083, young = 0.43, breast = 0.51 XXX = 1.0, XXX = 0.93, and so on), are assigned to fuzzification, the system then achieves the following MF values.

Medical

$= \{\mu$ (medicine) $= 1, \mu\,(XXX) = 1, \mu\,(breast)$
$= 0.21$ in adult interval and 0.91in medical interval,
$\mu\,(young) = 0.4$ in normal interval and 0.7
in medical interval}

Adult

$= \{\mu$ (babies) $= 0.8, \mu\,(XXX) = 1, \mu\,(breast)$
$= 0.21$ in adult interval and 0.91 in medical interval,
$\mu\,(girl) = 0.43$ in normal interval and 0.6
in adult interval}

The inference part applies the SWRL rule to the values of the fuzzy membership function as follows.

Medical content:

Rule 1 : if μ (medicine)
$= 1, \mu\,(XXX) = 1, \mu\,(breast) = 0.21$ or $0.91, \mu\,(young)$
$= 0.4$ or 0.7 then web type value is 0.5

Adult content:

Rule 1 : if μ (babies)
$= 0.8, \mu\,(XXX) = 1, \mu\,(breast) = 0.21$ or $0.91, \mu\,(girl)$
$= 0.43$ or 0.6 then web type value is 1

After extracting these rules, defuzzification converts the fuzzy output to crisp output, and offers the outcome in the form of a value, which is called an indicator value for webpage classification. The defuzzication is based on fitness function. The fitness function selects the MF degree for the antecedents of medical and adult content (words), and obtains the membership degree of consequent values as follows.

Medical content:

Fitness [1] $=$ min(μ (medicine)
$= 1, \mu\,(XXX) = 1, \mu\,(breast) = 0.21, \mu\,(young)$
$= 0.4) = 0.21$

Fitness [2] $=$ min(μ (medicine)
$= 1, \mu\,(XXX) = 1, \mu\,(breast) = 0.21, \mu\,(young)$
$= 0.7) = 0.21$

$$\begin{aligned}\text{Fitness}\,[3] &= \min(\mu\,(\text{medicine}) = 1, \mu\,(XXX)\\ &= 1, \mu\,(breast) = 0.91, \mu\,(young)\\ &= 0.4) = 0.41\\ \text{Fitness}\,[4] &= \min(\mu\,(\text{medicine}) = 1, \mu\,(XXX)\\ &= 1, \mu\,(breast) = 0.91, \mu\,(young)\\ &= 0.7) = 0.7\end{aligned}$$

Adult content:

$$\begin{aligned}\text{Fitness}\,[1] &= \min(\mu\,(\text{babies}) = 0.8, \mu\,(XXX)\\ &= 1, \mu\,(breast) = 0.21, \mu\,(girl)\\ &= 0.43) = 0.21\\ \text{Fitness}\,[2] &= \min(\mu\,(\text{babies}) = 0.8, \mu\,(XXX)\\ &= 1, \mu\,(breast) = 0.21, \mu\,(girl)\\ &= 0.6) = 0.21\\ \text{Fitness}\ [3] &= \min \mu\,(\text{babies}) = 0.8, \mu\,(XXX)\\ &= 1, \mu\,(breast) = 0.91, \mu\,(girl)\\ &= 0.43) = 0.43\\ \text{Fitness}\,[4] &= \min \mu\,(\text{babies}) = 0.8, \mu\,(XXX)\\ &= 1, \mu\,(breast) = 0.91, \mu\,(girl)\\ &= 0.6) = 0.6\end{aligned}$$

The final values for medical and adult content in the webpage are calculated using the following equations.

$$\text{Output}\,[i] = \text{fitness}\,[i] * \text{webtype}\,[i] \tag{3}$$

$$\text{Indicator value of webpage classification} = \frac{\sum \text{output}\,[i] * \text{fitness}\,[i]}{\sum \text{fitness}[i]} \tag{4}$$

Outputs 1, 2, 3, and 4 of the medical content are 0.105, 0.105, 0.2, and 0.35, respectively. Adult content outputs 1, 2, 3, and 4 are 0.21, 0.21, 0.43, and 0.6, respectively. The indicator values of webpage classification for the medical and adult content are 0.24 and 0.43, respectively. The sum of these outputs is 0.67. The intervals for each output are described as follows: the normal content interval is [0.0-0.4], medical content interval is [0.4-0.7], and adult content interval is [0.7-1]. Based on these intervals, the webpage classification value indicates the webpage is related to medical content.

V. EXPERIMENTS AND RESULTS

To verify the efficiency of the proposed system, a well-known method was used to evaluate the proposed fuzzy semantic knowledge in the form of queries and responses. A SPARQL query of Protégé OWL was applied to examine the performance of the ontology system. It is a target query to determine the relation between variables and properties. In addition, the reasoners gather the information based on defined rules in the fuzzy ontology. Therefore, it is important to execute the reasoner before any query execution. These reasoners are FACT++, DL reasoner, RacePro, and Pellet, which obtain the inference results from the ontology. However, we used the DL reasoner utility to evaluate the proposed ontology. It analyses the classical ontology and extracts the results. The proposed ontology has many imprecise terms, and sometimes, it is difficult to treat vague information in the web content. Therefore, we used the DL reasoner, because it is based on the Jena API and produces reasoning from the ontology [62]. It converts a fuzzy ontology into a crisp ontology and evaluates the performance of the ontology. Some SPARQL queries are considered to thoroughly analyze the overall effectiveness of the proposed system and retrieve individuals (instances/results) according to the requirements. These queries include the following.

SPARQL query to extract web content related to medical or normal webpage:

Syntax: *SELECT* WHERE {*
{?url url:HasTitle ?Has Title.
?url url:HasMetaData ?HasMetaData.
?url url:HasMetaKey ?HasMetaKey.}
}

Explanation: In this query, the ontology expert wants to extract the retrieved information related to medical and normal websites. These medical and normal classes hold all properties and their relationships with other classes to retrieve useful information. It has 1787 records along with URLs and meaningful keywords. Based on these classes, the system extracts the web content automatically before accessing the webpage, and saves it in the ontology for future use. The output of this query is shown in Fig. 6, which illustrates those URLs along with the web content that is allowed to be accessed with the proposed system.

SPARQL query to extract web content related to pornography:

Syntax: *SELECT* WHERE {*
{?subject rdfs:subClassOf ?object.
?url url:HasTitle ?Has Title.
?url url:HasMetaData ?HasMetaData.
FILTER REGEX (str(?object), 'Girls'), FITLER REGEX (str(?Has Title), 'sex').}
}

Explanation: The above SPARQL query filters the ontology to fetch the word *Girls* as an object class and *sex* as a word in the title tag. This query shows those URLs along with web content that contains pornographic words listed in the adult content. The adult content class has 2385 records along with information on URLs, valuable keywords, and other tags.

The performance of the adult content detection and classification system was evaluated by well-known methods, such as false-positive (FP) and false-negative (FN) ratings. We randomly collected 4646 webpages of the three different types. The system analyzed the content of these webpages and classified them into the three categories: adult content, medical, and normal webpages. Table 2 shows the dataset of content classification. The proposed system classified 1787 webpages as normal, 608 as medical, and 2251 as pornographic. For each webpage request, the lexicon dictionary, along with a set of features, was employed to calculate the indicator value for webpage classification and block access to adult content.

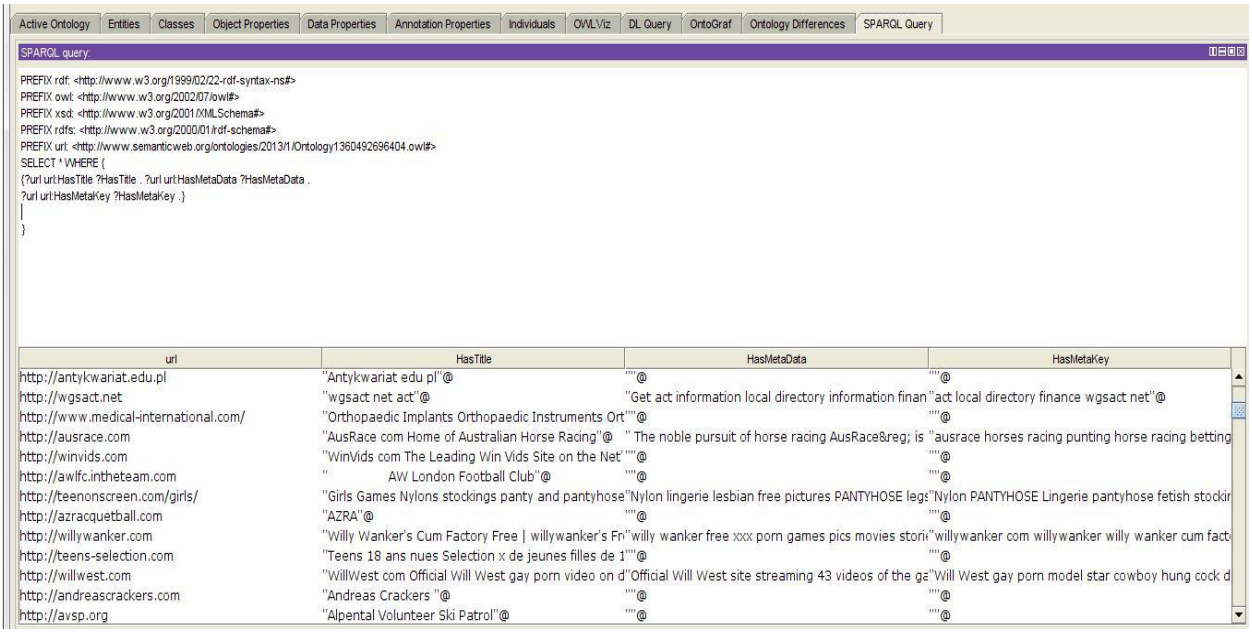

FIGURE 6. The output of a SPARQL query regarding medical and normal webpage data.

TABLE 2. The dataset of webpage classification.

Review	Webpages	Ratio
Normal content	1787	38.46%
Adult content	2251	48.45%
Medical content	608	13.08%
Total	4646	100%

This value classifies the webpage as normal if the value is below 0.4, as medical if the value is between 0.4 and 0.7, or as adult content if the value is above 0.7.

Three types of webpages were used to evaluate the adult content detection system. These webpages are related to medicine (like http://www.breastcancer.org), education (like http://antykwariat.edu.pl/), and pornography. The indicator values for education and medical webpages are 0.0468 and 0.67, respectively. These values show that if a webpage is related to the normal or medical class, the system allows users to access it. The calculation process of the indicator value for a medical webpage is explained in Section 4.3. On the other hand, an indicator value for an adult content webpage of 0.8247 illustrates that the webpage belongs in the pornography class, and the system denies access to this webpage automatically. Mathematically, the false-positive (FP) and false-negative (FN) rates can be calculated using the following equations:

$$\text{False Positive rate} = \frac{\text{NIP}}{\text{NCP} + \text{NIP} + \text{NIU}} \times 100\% \quad (5)$$

$$\text{False Negative rate} = \frac{\text{PIP}}{\text{PCP} + \text{PIP} + \text{PIU}} \times 100\% \quad (6)$$

In the equations above, NIP indicates the total number of negative instances incorrectly categorized as positive. Similarly, NCP, NIU, PIP, PCP, and PIU indicate the total number of negative instances that were correctly processed, the total number of negative instances that were incorrectly labeled unsure, the total number of positive instances that were incorrectly processed, the total number of positive instances that were correctly processed, and the total number of positive instances that were incorrectly labeled as unsure, respectively. The false-positive rate is proportional to the false-negative rate. The FP rate of webpages is shown in Table 3.The FP rate of negative-instance webpages (adult content) is 5.2%. This result shows that access to 94.8% of the webpages was disabled correctly, whereas access to 5.2% of the webpages was allowed by the proposed system because they contain other language words. Similarly, the FP rate of positive instance webpages (medical) is 0.9%. This result shows that the proposed system correctly allows access to 99.1% of the medical webpages, whereas access to 0.9% of the webpages was blocked due to strings of other language words. The FP rates of the proposed system were compared with existing static and dynamic filtering systems. Table 4 tabulates the FP rates of static filtering, dynamic filtering, and the proposed system. The FP rates of static and dynamic filtering listed in Table 4 indicate that existing static and dynamic filtering systems block 76.50% and 88.79% of the pornographic webpages, respectively, whereas the proposed system achieves a 94.8% FP rate in terms of adult content webpage blocking. It shows that the average improvement of the proposed system over the existing system is 18% in terms of FP rate. In addition, it is important to find the correct threshold (t) value for the adult content

TABLE 3. False-positive rate of webpages.

Webpages		Negative	Positive	Unsure	FP
Actual	Negative instances	TN = 1638	FP = 93	56	5.2 %
	Positive instances	FN = 28	TP = 2385	446	0.9%

TABLE 4. The FP rates of static filtering, dynamic filtering, and the proposed system.

System	False-Positive Rate
Static filtering system	23.50%
Dynamic filtering system	11.21%
The proposed system	5.2%

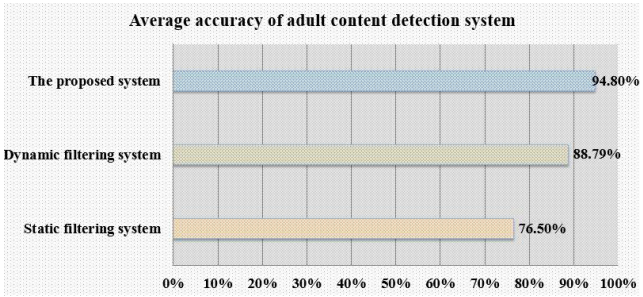

FIGURE 7. Graphical comparison of existing systems with the proposed system.

detection system. During computation of the FP rate, different threshold values were employed to evaluate the proposed mechanism. Table 5 presents the FP rate of the static filtering system, the dynamic filtering system, and the proposed adult content filtering system with different threshold values. The results for threshold t = 0.7 are compared with t = 0.4. It is obvious that the false-positive results at t = 0.7 are much better than at t = 0.4. It also shows that if the threshold value is decreased, then the system starts over when blocking webpages. The results for the proposed system show that access to adult content webpages can be prevented more accurately by using a fuzzy ontology and SVM with the threshold value t = 0.7. Fig. 7 clearly demonstrates the performance of the proposed system along with existing systems. It illustrates that the average accuracy of the proposed fuzzy ontology and SVM–based adult content detection system is better than existing static and dynamic filtering systems.

A. PERFORMANCE COMPARISON

In order to evaluate the performance of the proposed fuzzy ontology, prominent metrics of precision, recall, accuracy, and function measure are defined [54] and used to compare it with n-gram, KNN, and SVM classifiers.

$$\text{Precision (P)} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \times 100\% \tag{7}$$

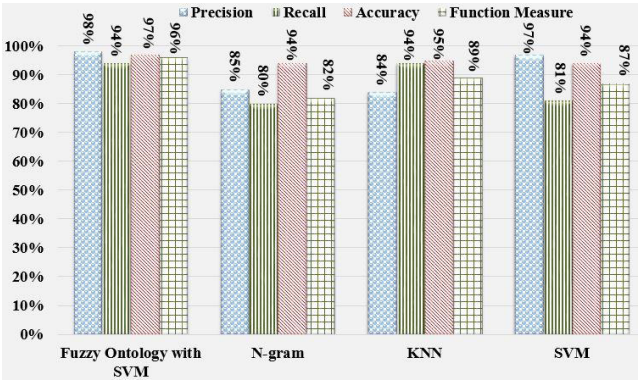

FIGURE 8. Graphical comparison among fuzzy ontologies with SVM, N-gram, KNN, and SVM.

$$\text{Recall (R)} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \times 100\% \tag{8}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \tag{9}$$

$$\text{Function Measure} = 2 * \frac{\text{P*R}}{\text{P} + \text{R}} \tag{10}$$

where TP, FP, FN, and TN, respectively, denote true positive, false positive, false negative, and true negative in adult content classification. The experimental results of the fuzzy ontology, n-gram, KNN, and SVM regarding adult content detection are shown in Table 6. It is obvious that the n-gram technique achieves good accuracy of 94%. However, the precision, recall, and function measure are just 85%, 80%, and 82%, respectively. On the other hand, the KNN technique obtains the best accuracy and recall at up to 95% and 94%, respectively. This technique however produces precision and function measure of 84% and 89%, respectively. With SVM, accuracy and precision are 95% and 97%, respectively, which are rather good results, but recall and function measure are just 81% and 87%, respectively.

In contrast, the proposed fuzzy ontology with SVM outperforms all classifiers in terms of all the metrics. The comparative results highlight the superiority of the proposed fuzzy ontology with an SVM-based system, compared to the n-gram, KNN, and SVM techniques in adult content detection and classification. Fig. 10 shows the performance comparisons among different classifiers using web content.

TABLE 5. The FP rate of static, dynamic, and the proposed system with different threshold values.

Threshold (t)	Static filtering	Dynamic filtering	Fuzzy ontology and SVM–based filtering	Total	FP rate
t=0.4	60.6%	14.33%	2.91%	77.84%	25.51
t=0.5	70.12%	15.5%	2.97%	88.59%	15.61
t=0.6	77.78%	16.62%	2.94%	97.34%	7.6
t=0.7	80.02%	16.84%	3.14%	100%	5.2

TABLE 6. Precision, recall, accuracy, and function measure for the fuzzy ontology, n-gram, KNN, and SVM methods.

Classification Method	Precision	Recall	Accuracy	Function Measure
Fuzzy Ontology with SVM	98%	94%	97%	96%
N-gram	85%	80%	94%	82%
KNN	84%	94%	95%	89%
SVM	97%	81%	94%	87%

VI. CONCLUSION

In this paper, the idea of a fuzzy-ontology/SVM–based adult content detection system is proposed to automate the classification of pornographic versus medical websites. Different sensibility issues are considered, including identification of the URL category, intelligent web content analysis to detect adult content, extraction of meaningful keywords, the use of unsupervised linear techniques to filter out irrelevant words, ontology-based semantic knowledge, and indicator value computation by using fuzzy logic to find the webpage type. The proposed mechanism offers an adult content detection system that classifies webpages into normal, pornographic, or medical webpages using extracted web content features. Indeed, the proposed mechanism successfully extracts useful keywords from web content and retrieves fuzzy variables to compute the indicator values for webpage classification. This technique can systematically retrieve all the web content and analyze this content to identify and block adult content websites. It can also be used at home, in offices and schools, and in other public sectors to intelligently investigate a network. Furthermore, this system can overcome the classification problem with medical websites, since it can extract medical features from unclear webpage content, classify these features as either medical or adult, and calculate an indicator value for the decision-making system. In future work, the detection method for adult content will be further improved by using a type-2 fuzzy neural network.

REFERENCES

[1] J. Wehrmann, G. S. Simões, R. C. Barros, and V. F. Cavalcante, "Adult content detection in videos with convolutional and recurrent neural networks," *Neurocomputing*, vol. 272, pp. 432–438, Jan. 2018.

[2] R. Cohen-Almagor, "Online child sex offenders: Challenges and countermeasures," *Howard J. Criminal Justice*, vol. 52, no. 2, pp. 190–215, 2013.

[3] E. Gatial, Z. Balogh, M. Laclavik, M. Ciglan, and L. Hluchy, "Focused Web crawling mechanism based on page relevance," in *Proc. ITAT*, 2005, pp. 41–46.

[4] J. J. Sheu, "Distinguishing medical Web pages from pornographic ones: An efficient pornography websites filtering method," *IJ Netw. Secur.*, vol. 19, no. 5, pp. 839–850, 2017.

[5] M. Bansal and J. Arora, "A novel OBIRS system for ontology based information retrieval system," *Int. J. Eng. Develop. Res.*, vol. 4, no. 2, pp. 1486–1489, 2016.

[6] T.-A. Dinh, T.-B. Ngo, and D.-L. Vu, "A model for automatically detecting and blocking pornographic websites," in *Proc. IEEE Int. Conf. Knowl. Syst. Eng. (KSE)*, Oct. 2015, pp. 244–249.

[7] J. Zhang and W.-Z. Ding, "An improved ontology-based Web information extraction," in *Proc. Int. Conf. Ed. Innov. Through Technol. (EITT)*, Oct. 2015, pp. 37–41.

[8] D. Sanchez, D. Isern, and A. Moreno, "Integrated agent-based approach for ontology-driven Web filtering," in *Knowledge-Based Intelligent Information and Engineering Systems*. Berlin, Germnay: Springer, 2006, pp. 758–765.

[9] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-based filtering in on-line social networks," in *Privacy and Security Issues in Data Mining and Machine Learning* (Lecture Notes in Computer Science). Berlin, Germnay: Springer, 2011, pp. 127–140.

[10] M. Wesam, A. Nabki, E. Fidalgo, E. Alegre, and I. De Paz, "Classifying illegal activities on Tor network based on Web textual contents," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 1. 2017, pp. 35–43.

[11] S. Seifollahi, I. Gondal, A. Bagirov, and R. Layton, "Optimization based clustering algorithms for authorship analysis of phishing emails," *Neural Process. Lett.*, vol. 46, no. 2, pp. 411–425, 2017.

[12] G. Xu, C. Wang, H. Yao, and Q. Qi, "Research on Tibetan hot words, sensitive words tracking and public opinion classification," *Cluster Comput.*, 2017. [Online]. Available: https://doi.org/10.1007/s10586-017-1026-x

[13] D. Yu, N. Chen, F. Jiang, B. Fu, and A. Qin, "Constrained NMF-based semi-supervised learning for social media spammer detection," *J. Knowl.-Based Syst.*, vol. 125, pp. 64–73, Jun. 2017.

[14] Z. Weinberg, M. Sharif, J. Szurdi, and N. Christin, "Topics of controversy: An empirical analysis of Web censorship lists," *Proc. Privacy Enhancing Technol.*, vol. 217, no. 1, pp. 42–61, 2017.

[15] F. Salvetti and N. Nicolov, "Weblog classification for fast Splog filtering: A URL language model segmentation approach," in *Proc. Human Lang. Technol. Conf. (NAACL)*, Jun. 2006, pp. 137–140.

[16] E. Baykan, M. Henzinger, L. Marian, and I. Weber, "Purely URL-based topic classification," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 1109–1110.

[17] I. Hernández, C. R. Rivero, D. Ruiz, and R. Corchuelo, "A statistical approach to URL-based Web page clustering," in *Proc. 21st Int. Conf. Companion World Wide Web*, 2012, p. 525.

[18] C. J. Dietrich and C. Rossow, "Empirical research of IP blacklists," in *Securing Electronic Business Processes*. Wiesbaden, Germany: Vieweg+Teubner, 2008, p. 8.

[19] Y.-H. Feng, N.-F. Huang, and C.-H. Chen, "An efficient caching mechanism for network-based URL filtering by multi-level counting bloom filters," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2011, pp. 1–6.
[20] L.-H. Lee and H.-H. Chen, "Collaborative cyberporn filtering with collective intelligence," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 1153–1154.
[21] J. M. G. Hidalgo, F. C. García, and E. P. Sanz, "Named entity recognition for Web content filtering," in *Natural Language Processing and Information Systems*. Berlin, Germnay: Springer, 2005, pp. 286–297.
[22] H. Ma, "Fast blocking of undesirable Web pages on client PC by discriminating URL using neural networks," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 1533–1540, 2008.
[23] G. M. Thomaz, A. A. Biz, E. M. Bettoni, L. Mendes-Filho, and D. Buhalis, "Content mining framework in social media: A FIFA world cup 2014 case analysis," *Inf. Manage.*, vol. 54, no. 6, pp. 786–801, 2017.
[24] A. Kova, "Cyberbullying detection using Web content," in *Proc. 22nd Telecommun. Forum Telfor (TELFOR)*, 2014, pp. 939–942.
[25] Z. Gao, G. Lu, H. Dong, S. Wang, H. Wang, and X. Wei, "Applying a novel combined classifier for hypertext classification in pornographic Web filtering," in *Proc. Int. Conf. Internet Comput. Sci. Eng. (ICICSE)*, 2008, pp. 270–273.
[26] V. F. FernÃąndez, R. M. Unanue, S. M. Herranz, and A. C. Rubio, "Naïve Bayes Web page classification with HTML mark-up enrichment,," in *Proc. Int. Multi-Conf. Comput. Global Inf. Technol. (ICCGI)*, vol. 6. 2007, p. 48.
[27] S. Sinha, M. Bailey, F. Jahanian, and A. Arbor, "Improving spam blacklisting through dynamic thresholding and speculative aggregation," in *Proc. 17th Annu. Netw. Distrib. Syst. Secur. Symp.*, 2010, pp. 1–2.
[28] W. K. Chan, Y. Y. Chiu, and Y. T. Yu, "A Web search-centric approach to recommender systems with URLs as minimal user contexts," *J. Syst. Softw.*, vol. 84, no. 6, pp. 930–941, 2011.
[29] A. C. M. Fong, S. C. Hui, P. Y. Lee, M. Hammami, R. Guermazi, and B. Hamadou, "XFighter: An intelligent Web content filtering system," *Int. J. Web Inf. Syst.*, vol. 38, no. 4, pp. 1541–1555, 2009.
[30] W. H. Ho and P. A. Watters, "Identifying and blocking pornographic content," in *Proc.-Int. Workshop Biomed. Data Eng.*, 2005, p. 1181.
[31] H. Zheng, H. Liu, and M. Daoudi, "Blocking objectionable images: Adult images and harmful symbols," in *Proc. ICME*, 2004, pp. 2–5.
[32] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank, "Recognition of pornographic Web pages by classifying texts and images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1019–1034, Jun. 2007.
[33] C. Fan, J. Song, Z. Wen, and Y. Wu, "Content semantic filter based on domain ontology," in *Proc. IEEE Int. Conf. Progr. Informat. Comput. (PIC)*, vol. 1. Dec. 2010, pp. 234–237.
[34] M. Vanzin, K. Becker, and D. D. A. Ruiz, "Ontology-based filtering mechanisms for Web usage patterns retrieval," in *Proc. Int. Conf. Electron. Commerce Web Technol.*, 2005, pp. 267–277.
[35] M. Fern, D. Vallet, and P. Castells, "A multi-purpose ontology-based approach for personalised content filtering," *Adv. Semantic*, vol. 51, no. 2008, pp. 25–51, 2008.
[36] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious Web sites from suspicious URLs," in *Proc. World Wide Web Internet Web Inf. Syst.*, 2009, pp. 1245–1253.
[37] G. Su, J. Li, Y. Ma, and S. Li, "Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model," *J. Zhejiang Univ. Sci.*, vol. 5, no. 9, pp. 1106–1113, 2004.
[38] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler?: A fast filter for the large-scale detection of malicious Web pages categories and subject descriptors," in *Proc. Int. World Wide Web Conf. (WWW)*, 2011, pp. 197–206.
[39] P.-C. Lin, M.-D. Liu, Y.-D. Lin, and Y.-C. Lai, "An early decision algorithm to accelerate Web content filtering," in *Information Networking. Advances in Data Communications and Wireless Networks*. Berlin, Germany: Springer-Verlag, 2006, pp. 833–841.
[40] T. Fu and H. Chen, "Analysis of cyberactivism: A case study of online free Tibet activities," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.*, 2008, pp. 1–6.
[41] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, M. D. M. Coelho, and A. D. A. Araújo, "Nude detection in video using bag-of-visual-features," in *Proc. SIBGRAPI*, 2009, pp. 224–231.
[42] A. Behrad, M. Salehpour, M. Saiedi, and M. N. Barati, "Obscene video recognition using fuzzy SVM and new sets of features," *Int. J. Adv. Robot. Syst.*, vol. 10, no. 2, pp. 1–11, 2013.
[43] Z. Zhou, T. Song, and Y. Jia, "A high-performance URL lookup engine for URL filtering systems," in *Proc. IEEE Int. Conf. Commun.*, May 2010, pp. 1–5.
[44] C. S. Lee, M. H. Wang, M. H. Wu, C. Y. Hsu, Y. C. Lin, and S. J. Yen, "A type-2 fuzzy personal ontology for meeting scheduling system," in *Proc. IEEE World Congr. Comput. Intell. (WCCI)*, Jul. 2010, pp. 1–8.
[45] P. Yan, Y. Zhao, and C. Sanxing, "Ontology-based information content security analysis," in *Proc.-5th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, 2008, pp. 479–483.
[46] F. Ali, E. K. Kim, and Y. G. Kim, "Type-2 fuzzy ontology-based opinion mining and information extraction: A proposal to automate the hotel reservation system," *Appl. Intell.*, vol. 42, no. 3, pp. 481–500, 2015.
[47] S. Nasrolahi, M. Nikdast, and M. M. Boroujerdi, "The semantic Web: A new approach for future," *Semantic Web, Approach Future World Wide Web*, vol. 34, pp. 1149–1154, Mar. 2009.
[48] A. Paar, J. Reuter, J. Soldatos, K. Stamatis, and L. Polymenakos, "A formally specified ontology management API as a registry for ubiquitous computing systems," *Appl. Intell.*, vol. 30, no. 1, pp. 37–46, 2009.
[49] B. Shah, F. Iqbal, A. Abbas, and K.-I. Kim, "Fuzzy logic-based guaranteed lifetime protocol for real-time wireless sensor networks," *Sensors*, vol. 15, no. 8, pp. 20373–20391, 2015.
[50] B. Shah, A. M. Khattak, and K.-I. Kim, "A fuzzy logic scheme for real-time routing in wireless sensor networks," in *Proc. 6th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, 2015, pp. 1–7.
[51] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965.
[52] Z. S. Mi, A. C. Bukhari, and Y. G. Kim, "An obstacle recognizing mechanism for autonomous underwater vehicles powered by fuzzy domain ontology and support vector machine," *Math. Problems Eng.*, vol. 2014, Aug. 2014, Art. no. 676729.
[53] F. Ali *et al.*, "Merged ontology and SVM-based information extraction and recommendation system for social robots," *IEEE Access*, vol. 5, pp. 1–16, 2017.
[54] F. Ali, K.-S. Kwak, and Y.-G. Kim, "Opinion mining based on fuzzy domain ontology and support vector machine: A proposal to automate online review classification," *Appl. Soft Comput.*, vol. 47, pp. 235–250, Oct. 2016.
[55] N. F. Noy, D. L. McGuinness. (Mar. 2001). "Ontology development 101: A guide to creating your first ontology." Tech. Rep. [online] Available: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf
[56] D. Lembo, M. Lenzerini, R. Rosati, M. Ruzzi, and D. F. Savo, "Inconsistency-tolerant query answering in ontology-based data access," *Web Semantics, Sci., Services Agents World Wide Web*, vol. 33, pp. 3–29, Aug. 2015.
[57] D. Calvanese *et al.*, "Ontop: Answering SPARQL queries over relational databases," *Semantic Web*, vol. 8, no. 3, pp. 471–487, 2016.
[58] T. Pixley, *Document Object Model (DOM) level 2 Events Specification*, W3C Recommendation, Nov. 2000. [online] Available: http://www.w3.org/tr/dom-level-2-events/
[59] F. Ali, D. Kwak, P. Khan, S. M. R. Islam, K. H. Kim, and K. S. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling," *Transp. Res. C, Emerg. Technol.*, vol. 77, pp. 33–48, Apr. 2017.
[60] F. Bobillo and U. Straccia, "The fuzzy ontology reasoner *fuzzyDL*," *Knowl.-Based Syst.*, vol. 95, pp. 12–34, Mar. 2016.
[61] S. El-Sappagh, M. Elmogy, and A. M. Riad, "A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis," *Artif. Intell. Med.*, vol. 65, no. 3, pp. 179–208, 2015.
[62] Q. Guo and M. Zhang, "A novel approach for multi-agent-based intelligent manufacturing system," *Inf. Sci.*, vol. 179, no. 18, pp. 3079–3090, 2009.

FARMAN ALI received the B.S. degree in computer science from the University of Peshawar, Pakistan, in 2011, the master's degree in computer science from Gyeongsang National University, South Korea, in 2015. He is currently pursuing the Ph.D. degree with the Department of Information and Communication Engineering, Inha University, South Korea. His current research interests include opinion mining, ontology, fuzzy logic, artificial intelligence, and machine learning.

PERVEZ KHAN received the bachelor's and master's degrees in computer science from the University of Peshawar, Pakistan, in 2006 and 2003, respectively, and the Ph.D. degree in information and communication engineering from the Graduate School of IT and Telecommunication Engineering, Inha University, Incheon, South Korea, in 2015. His current research interests include wireless communications, fuzzy logic, semantic analysis, wireless sensor networks, wireless ad-hoc networks, wireless body area networks, performance evaluation, and MAC protocol design.

KASHIF RIAZ received the B.S degree in computer science from the University of Punjab, Gujranwala Campus, Pakistan, in 2010, and the master's degree in computer science from the COMSATS Institute of Information Technology, Islamabad Campus, Pakistan, in 2014. He is currently a Lecturer with the Department of Information Technology, University of the Punjab, Gujranwala Campus, Pakistan. His current research area includes intelligent system, database system, information extraction, and soft computing.

DAEHAN KWAK received the B.S. degree in information and computer engineering from Ajou University, and the M.S. degree from the Korea Advanced Institute of Science and Technology, South Korea, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from Rutgers University, NJ, USA, in 2017. He is currently an Assistant Professor with Kean University, NJ, USA. He was with the Telematics and USN Research Division, Electronics and Telecommunications Research Institute. He had been with the UWB Wireless Research Center, Inha University, South Korea, as a Researcher. His current research interest includes mobile and pervasive computing, vehicular computing and networks, u-health networks/applications, and mobile applications.

TAMER ABUHMED received the Ph.D. degree in information and telecommunication engineering from Inha University in 2012. He is currently an Assistant Professor with the Department of Computer Engineering, Inha University, South Korea. His research interests include applied cryptography and information security, network security, Internet security, and machine learning and its application to security and privacy problems.

DAEYOUNG PARK received the B.S. and M.E. degrees in electrical engineering and the Ph.D. degree in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 1998, 2000, and 2004, respectively. He was with Samsung Electronics as a Senior Engineer from 2004 to 2007, contributing to the development of next-generation wireless systems based on MIMO-OFDM systems. He has held visiting positions at the University of Southern California, Los Angeles, CA, USA, and the University of California, San Diego, CA, USA. Since 2008, he has been with Inha University, Incheon, South Korea, where he is currently an Associate Professor. He is a co-author of *Wireless Communications Resource Management* (Wiley, 2009). His research interests include communication systems, wireless networks, multiuser information theory, and resource allocation.

KYUNG-SUP KWAK (M'81) received the Ph.D. degree from the University of California at San Diego in 1988. From 1988 to 1989, he was with Hughes Network Systems, San Diego, CA, USA. From 1989 to 1990, he was with the IBM Network Analysis Center, Research Triangle Park, NC, USA. Since then, he has been with the School of Information and Communication Engineering, Inha University, South Korea, as a Professor, where he had been the Dean of the Graduate School of Information Technology and Telecommunications from 2001 to 2002. He has been the Director of the UWB Wireless Communications Research Center, South Korea, since 2003. In 2006, he served as the President of Korean Institute of Communication Sciences, and in 2009, the President of the Korea Institute of Intelligent Transport Systems. In 2008, he had been selected for Inha Fellow Professor and currently a Inha Hanlim Fellow Professor. He has authored over 200 peer-reviewed journal papers and served as TPC, the Track chairs, and the Organizing Chairs for several IEEE related conferences. His research interests include wireless communications, UWB systems, sensor networks, WBAN, and nano communications. He was a recipient of the number of awards, including the Engineering College Achievement Award from Inha University, the LG Paper Award, the Motorola Paper Award, the Haedong Prize of research, and various government awards from the Ministry of ICT, the President, and the Prime Minister of South Korea, for his excellent research performances.