IEEE
INTERNET OF THINGS JOURNAL
IEEE Xplore® Digital Library
http://ieeexplore.ieee.org/Xplore

Automatic Knowledge Extraction to build Semantic Web of Things Applications

Mahda Noura[§], Amelie Gyrard*[†], Sebastian Heil[§], and Martin Gaedke[§]
[§]Technische Universität Chemnitz, Germany
[†]Kno.e.sis, Wright State University, USA

Abstract—**The Internet of Things (IoT) primary objective is to make a hyper-connected world for various application domains. However, IoT suffers from a lack of interoperability leading to a substantial threat to the predicted economic value. Schema.org provides semantic interoperability to structure heterogeneous data on the Web. An extension of this vocabulary for the IoT domain (iot.schema.org) is an ongoing research effort to address semantic interoperability for the Web of Things (WoT). To design this vocabulary, a central challenge is to identify the main topics (concepts and properties) automatically from existing knowledge in IoT applications. We designed KE4WoT (Knowledge Extraction for the Web of Things) to automatically identify the most important topics from literature ontologies of 3 different IoT application domains – smart home, smart city and smart weather – based on our corpus consisting of 4500 full-text conference and journal articles to utilize domain-specific knowledge encoded within IoT publications. Despite the importance of automatically identifying the relevant topics for iot.schema.org, up to know there is no study dealing with this issue. To evaluate the extracted topics, we compare the descriptiveness of these topics for the 10 most popular ontologies in the 3 domains with empirical evaluations of 23 domain experts. The results illustrate that the identified main topics of IoT ontologies can be used to sufficiently describe existing ontologies as keywords.**

Index Terms—**Internet of Things (IoT), Web of Things (WoT), Knowledge Extraction, Machine Learning (ML), Natural Language Processing (NLP), Ontologies, Semantic Web of Things (SWoT)**

I. Introduction

The Internet of Things (IoT)'s vision is to connect all *Things* (Radio-Frequency IDentification, sensors, actuators, etc.) to the Internet, allowing a wide range of innovations and opportunities in different application domains [1]. Although there has been a massive growth in this domain, currently "developing a single and global ecosystem of Things that communicate with each other seamlessly is virtually impossible" [2]. To achieve the vision of IoT, in which "people and things connected anytime, anyplace, with anything and anyone" [3], interoperability at different layers is required [4], [5].

The W3C Web of Things (WoT) Initiative[1] intends to provide an interoperable infrastructure to simplify access to smart devices by making them controllable via the existing Web standards. Semantic interoperability describes smart devices according to their data, services, and capabilities in machine-readable form using a shared vocabulary. Unfortunately, the wide range of ontologies to represent IoT devices and their produced data hinders the efficient development of cross-platform and cross-domain applications [6], [7], [8]. Our analysis in [9] demonstrated that many of the ontologies found in existing standardizations and different projects have many redundant concepts and properties re-designed instead of reusing existing ones. It is time-consuming and challenging to systematically identify the most relevant IoT topics (concepts and properties) for reuse in new WoT applications.

Schema.org provides a semantic schema to describe websites to search engines explicitly. Companies such as Google design schema.org to structure data on the Web. Schema.org is extended to iot.schema.org, so WoT applications interact with the physical world. Schemas model IoT data and deduce meaningful information to build smarter WoT applications.

The development of iot.schema.org requires the identification of the most relevant topics in the IoT domain. This can be achieved by automatically analyzing the most relevant ones from existing IoT ontologies. Designing schemas require domain experts to follow a systematic approach and should not be based on a arbitrary manual process dependent on human preferences/judgment to agree on all the concepts and properties to include. It is a non-trivial problem which is time-consuming and requires several inter-disciplinary experts to understand the meaning of the different topics. The automatic analysis of topics helps domain experts to unify and design new ontologies.

To help domain experts, we design the novel **KE4WoT (Knowledge Extraction for Web of Things)** methodology which automatically analyzes the key topics that frequently appear in existing ontologies from a specific IoT application domain. KE4WoT employs machine learning techniques and relies on a systematically created corpus of 4500 full-text scientific articles in 3 different IoT application domains: home, city, and weather. The results of KE4WoT (key topics) especially helps the creators of iot.schema.org. The main advantage of KE4WoT is reducing the development time and the overall human workload for creating any schema, and increase topic re-use for better interoperability. Domain experts can then include the topics into a new vocabulary to describe IoT devices, services and capabilities to enable semantic interoperability for building composite applications across diverse WoT ecosystems.

Knowledge extraction from unstructured or semi-structured sources has seen significant attention [10] such as analyzing social network microblogs and posts (e.g., Twitter, Reddit)

* This work is partially funded by Hazards SEES NSF Award EAR 1520870, and KHealth NIH 1 R01 HD087132-01. The opinions expressed are those of the authors and do not reflect those of the sponsors.

[1] https://www.w3.org/WoT/

Digital Object Identifier: 10.1109/JIOT.2019.2918327

and examining text documents. However, knowledge extraction from structured knowledge sources (e.g., ontologies) has received scant attention. To the best of our knowledge, no previous work has proposed such a methodology incorporating machine learning (Word2vec and k-means clustering algorithms) techniques to analyze a set of IoT related ontologies.

Challenges: Long-term challenges to address are: (1) Designing a unified schema in a specific domain (e.g., a standard schema to describe units), (2) analyzing the most common topics of ontologies in a specific domain, and (3) classifying the topics. Designing a unified schema is a cornerstone component which:

- Classifies sensor types, and IoT application domains.
- Is an exploratory analysis[2] to contribute towards building iot.schema.org.
- Improves semantic interoperability among IoT projects when they adopt the schema.
- Classifies and integrates existing ontologies based on the scientific literature analysis.
- Encourages the reuse of domain knowledge expertise, though ontology reuse.

Contributions: The KE4WoT methodology extends our previous work ("Oustanding Paper Award" [9] and KE4WoT challenge[3]) by: 1) applying the methodology to 3 new IoT application domains: home, city, and weather (previously only covering generic IoT/WoT ontologies), 2) increasing the ontologies used for the process of extracting the main topics from 14 to 46, 3) creating a large IoT corpus by following a systematic approach (instead of using the pre-trained model from the Google News which does not include IoT/WoT terms). 4) demonstrating the applicability of the approach with three motivational scenarios: designing the taxonomy of sensors and IoT application domains, selecting the appropriate ontologies to enrich IoT data, and recommendations for W3C endorsed ontologies or schema.org, 5) our results[4] have been communicated to iot.schema.org, and 6) providing a detailed empirical evaluation of the descriptiveness of the identified concepts.

The remainder of the paper is structured as follows: Section II demonstrates three motivational scenarios. Section III introduces the KE4WoT methodology to extract "common sense knowledge" from ontologies. Section IV details implementation and results and in Section V our methodology is evaluated. Section VI summarizes the limitations of the existing literature study. Finally, Section VII concludes the paper and provides future insights.

II. Motivational Scenarios

The IoT ontological statistical analysis is motivated with the following three motivational scenarios: (1) Classifying sensors, and IoT application domains within the IoT dictionary, (2) interpreting IoT data, and (3) unifying ontologies.

(1) IoT Dictionary (Sensors and IoT Application Domains): The automatic analysis of new sensor types, devices and domains from existing ontologies enriches the IoT dictionary. A semi-automatic approach adds a new `ssn:Sensor` subclass from the W3C recommendation SSN/SOSA ontology[5]. The SSN ontology documentation[6] clearly demonstrates the usage of the SSN ontology V1 [11] (published in 2011) within sensor datasets and ontologies. For each ontology, a table shows whether SSN modules (e.g., `Observations`, `FeatureofInterest`), concept (e.g, `Sensor`), and properties (e.g., `observes`) are employed or not. The documentation has been done manually which demonstrates the need of an automatic approach, generic enough to be executed on any IoT domains. The SSN/SOSA missing pieces[7] are: (1) Standardized taxonomy of observable properties (e.g., temperature), (2) The controlled vocabulary of feature of interests per domain (e.g., air), (3) Industry controlled vocabulary of sensor/actuator types (e.g., thermometer), and (4) Standardized industry procedures (e.g., on/off).

The M3 ontology [12] has been manually built based on a state of the art analysis and covers limitations of the W3C SSN ontology. The M3 ontology is a dictionary[8] which provides: 1) sensor type list, 2) observable property list, 3) feature of interest list, and 4) unit list. The m3-lite extension, integrated within the FIESTA-IoT ontology [13], unifies a set of IoT ontologies. This tedious manual task of unifying and reusing ontologies to enhance IoT semantic interoperability highlights the need to automate this process: 1) to maintain and enrich the ontology, 2) to quickly reproduce the alignment process in other topics. As a future work, the WordNet "synset" synonym dictionary can automate better this task.

(2) Interpreting IoT Data: Selecting the appropriate designed ontologies is required to process and interpret IoT data according to the application's specifications (e.g., the m3-lite ontology is employed with machine learning algorithms for road and traffic analysis [14]. Sometimes IF THEN ELSE RULES are defined as `owl:Restriction` within ontologies (e.g., Staroch's ontology [15] provides simple rules such as if precipitation = 0 mm then no precipitation and more complex rules.

(3) Unifying Schemas: W3C endorsed ontologies and iot.schema.org demonstrate the need to unify IoT schemas. The W3C SSN ontology has been manually designed and combines concepts from several sensor ontologies (explained above). Automating the analysis of existing ontologies to detect common patterns such as a list of sensors used, sensor measurements types, and its IoT application domains reduce the human workload.

III. Knowledge Extraction for the Web of Things: KE4WoT Methodology

The automatic knowledge extraction KE4WoT methodology (depicted in Figure 1) defines two roles: 1) the *Expert* conducts the knowledge extraction process, and 2) the *Analysis Toolchain*, a system that supports the knowledge extraction

[2] https://goo.gl/BPi17x

[3] http://wiki.knoesis.org/index.php/KE4WoTChallengeWWW2018

[4] Slides: https://goo.gl/92RKCb

[5] https://www.w3.org/TR/vocab-ssn/

[6] https://w3c.github.io/sdw/ssn-usage/

[7] Armin Haller, co-editor of SSN/SOSA, concluded his talk at the ISWC SSN 2018 workshop https://goo.gl/x5Wzwn

[8] http://sensormeasurement.appspot.com/?p=m3

process using code scripts. A challenge is the development of a corpus for training the word2vec algorithm to comprehend the meaning of the different concepts and properties defined in the IoT ontologies; such IoT corpus is missing in the literature. Section III-A describes the KE4WoT conceptual process, and Section III-B explains the creation of the IoT corpus.

A. KE4WoT Conceptual Process

The KE4WoT conceptual process comprises 10 steps:

Step 1 - Ontology Selection: the *expert* selects the ontologies (in RDF, RDFS, and OWL), from the LOV4IoT catalog [16], [17], to be included for the knowledge extraction process. LOV4IoT provides almost 480 IoT-based ontologies for 21 domains. The ontology selection criteria is detailed in Section V-A).

Step 2 - Pre-processing: the cleaned ontologies are stored in a Virtuoso triplestore. The experts convert: 1) all the ontologies into TTL representation using Protege, and 2) unicode-encoded characters into ASCII since many commonly used ontology parsers do not handle Unicode-encoded text properly.

Step 3 - Vocabulary Extraction: the *analysis toolchain* queries ontology terms such as classes, subclasses, properties, labels, and SKOS concepts expressed in RDF form. The output is a list of unique vocabularies.

Step 4 - Term Extraction: the set of terms are extracted from an ontology as follows:

1) Splitting multi-words based on snake-case and camel-case to create a list of unique single words (e.g., splitting "MultiDevice" to two words "multi" and "device").
2) Removing special characters (e.g., punctuation marks, numbers, whitespaces, etc.) to prevent terms such as "error24." and "error." indexed as two different terms.
3) Removal of stop words which do not add any useful information (i.e., prepositions, connecting words, conjunctions, etc.).
4) Converting all the terms to lower case (e.g., "IOT/iot").

Step 5 - Frequency Calculation: the frequency of each term in all ontologies is computed and used as a metric for identifying the most important topics.

Step 6 - Word2vec algorithm: identifies the association of a word with other words. The Google word2vec algorithm is a neural network used to process and convert text into a set of numerical vectors. The vectors represent the distance between individual terms. word2vec makes accurate guesses about the meaning of a word (if enough data and context). The word2vec algorithm requires learning the vectors. We trained word2vec by creating a corpus from IoT scientific publications (explained in Section III-B). The trained model is run on the unique terms (from the previous step) to provide the word embeddings as output. To reduce the dimension of unique words, the *min-count* (minimum frequency) variable in the word2vec algorithm is set to 5 to ignore words that are infrequent.

Step 7 - K-means clustering: semantically coherent terms are clustered together by exploiting the word2vec word embeddings using the unsupervised k-means machine learning algorithm. The K parameter is defined as the number of clusters (i.e., a set of ontology terms which represents semantically coherent terms). K parameters for each domain are provided in Section IV.

TABLE I: Number of retrieved scientific publications per IoT application domain.

Application Domain	Number of articles
Smart Home	1162
Smart City	1322
Smart Weather	2019

Step 8 - Results Aggregation: clusters are identified: 1) the sum of term occurrences in all ontologies, and 2) the ontology list that include the cluster terms.

Step 9 - Topic Name Assignment: the *expert* manually assigns the cluster name to the term list, defined as the topic.

Step 10 - Popular Topic Identification: two results are produced: 1) the list of ordered terms based on the frequency, and 2) the ontology list when terms identified are used within the ontology.

B. Creation of the IoT Text Corpus

To train the word2vec algorithm (see Figure 1), we created the IoT text Corpus for understanding similar terms in the ontology. There is no pre-trained corpus in the literature covering the different IoT/WoT domains. The corpus creation involves two main steps: 1) Corpus Creation, and (2) Corpus Processing.

Step 1 - Corpus Creation of three sub-corpora containing scientific articles from 3 IoT application domains (home, city, and weather forecast). The corpus creation follows Kitchenham [18] guidelines. We performed a combination of automated and manual searches of scientific literature for domains that we cover.

The search process queries computer science scientific libraries (Google Scholar, ACM Digital Library, Springer, Science Direct, and CiteSeer) to cover a wide range of vocabularies in a specific IoT domain. Searches are performed using advanced search query strings to analyze the scientific publications (entire text, keyword, title, and abstract). The search query has two parts: the synonyms for IoT and the application domain (e.g., smart home, smart city or weather). An example of queries: `("internet of things" OR "IoT" OR "smart object" OR "web of things") AND (<application domain>)`

Table I references the number of scientific publications per application domain, published between 2012 and 2018.

Step 2 - Corpus Processing analyzes the texts gathered from the selected publications before using them with word2vec. The unprocessed corpus contains a variety of basic syntactic variations and punctuation that degrades the performance of the word2vec algorithm for vector models. The processing of each corpus involved the following main steps: 1) converting all downloaded articles from PDF to text format, 2) removing all punctuation signs, numbers, and extra whitespace, 3) removal of stop words, 4) transforming all words to lower-case, and 5) and removing the reference section of the articles.

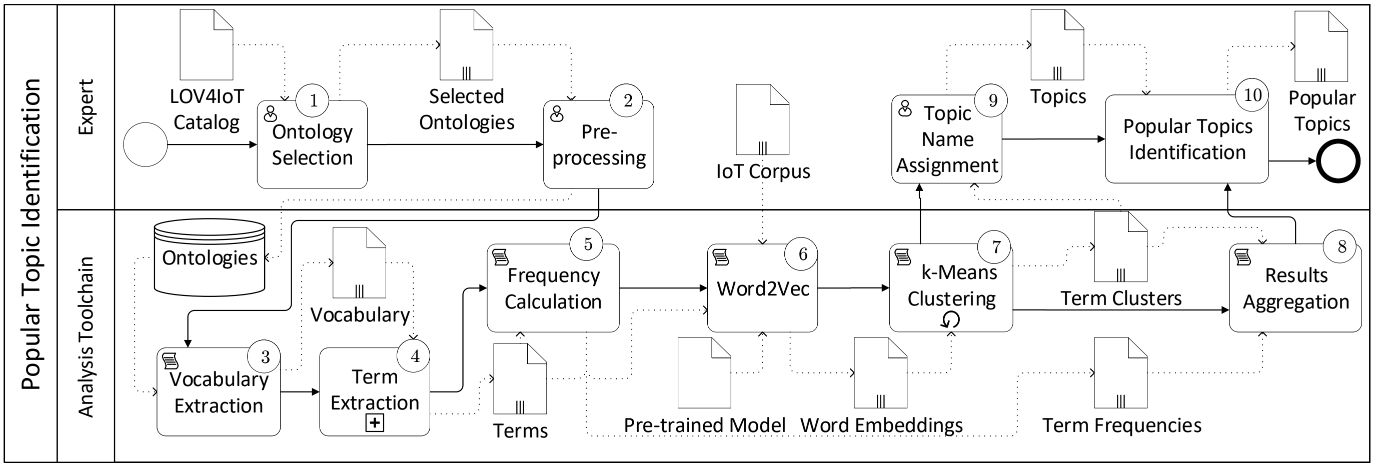

Fig. 1: The KE4WoT Methodology extracts common concepts and properties from ontologies

IV. Implementation and Results

KE4WoT is implemented with python scripts: 1) the scrapy library[9] crawls the articles from scientific libraries mentioned above, 2) the ontospy[10] parses the ontologies, and 3) the NLTK library[11] pre-processes the text from articles for the word2vec model and k-means algorithm. The clustering experiments is performed with the k-means algorithm with different values of k to achieve semantically consistent clusters for each domain. The value of the K for each domain is determined by visualizing the keywords between the clusters to identify how the keywords are split. To assign names to the clusters resulting from step 9, the terms within each cluster were reviewed by three researchers and based on the commonalities of the terms, a categorial name is selected. The naming process used additional hypernym information derived from WebIsALOD[12] the Linked Open Data version of the WebIsA Database, a database containing 11.7 million hypernymy relations extracted from the CommonCrawl web corpus. For that, we ran hypernym queries on the SPARQL endpoint of WebIsALOD for each cluster member and merged the result sets to identify the most common hypernyms of all terms in the cluster. Since the resulting most frequent hypernyms tend to be very general concepts (e.g., information, thing, business), the naming process could not be fully automated, but required human assessment informed by the hypernym query results.

The rest of this section reports the identification of the most frequent terms for smart home (Section IV-A1), weather (Section IV-A2), and city (Section IV-A3). The discussion is provided in Section IV-B.

A. Extracting Knowledge from 3 domains

The complete cluster results, per application domain, are available online[13]. For brevity, we list the results of the 15 most frequent topics for home (Table II), city (Table IV) and weather (Table III). The first column is the identified name cluster. The second column contains some example terms (and not the complete set of all terms in that cluster). In column three, the aggregated term frequencies of all terms of the topic among different ontologies is provided. The last column provides the number of ontologies that use the members of the topic.

1) Extracting Smart Home Knowledge: The LOV4IoT smart home dataset comprising **24 smart home ontologies** has been analyzed. The value of k=87 produces the most meaningful and semantically consistent clusters for this domain. Table II shows the results of the 15 most popular terms within home ontologies. Each cluster demonstrates a set of semantically related terms. For instance, the *Appliance* cluster name indicates the required appliances to build a smart home (e.g., sensor, device, smartphone, etc.). The *Home safety & security* cluster name is stressed more than devices in the current home ontologies, despite the high diversity of smart home devices in the market.

2) Extracting Weather Forecasting Knowledge: Table III shows the results of the KE4WoT methodology executed on the LOV4IoT weather dataset comprising **10 ontologies**. The value of k=30 produces the most meaningful and semantically consistent clusters with regards to the weather domain. As expected, the most frequent terms in the weather domain are weather phenomena which is found in all weather ontologies.

3) Extracting Smart City Knowledge: Table IV shows the most important terms identified with the KE4WoT methodology on the LOV4IoT smart city dataset comprising **12 ontologies**. The dataset includes impactful ontologies such as STAR-CITY [19] (developed by IBM) which has been deployed in four cities (Dublin, Bologna, Miami, Rio de Janeiro). The k-mean algorithm is applied with the value k=82 since it represents semantically consistent clusters for the smart city domain. The *Smart city topics* cluster name clearly demonstrates the different topics related to this domain (e.g., mobility, tourism, grid, traffic, etc.). The results indicate that *Country statistical profile* is the most important term in the smart city domain, where only one ontology ([20]) uses it. It is mainly because the Komninos ontology is the largest ontology

[9] https://scrapy.org/
[10] https://pypi.org/project/ontospy/
[11] https://www.nltk.org/
[12] http://webisa.webdatacommons.org
[13] https://goo.gl/y9VwoU

TABLE II: The most important smart home terms among 24 different ontologies.
Legend: F=Frequency, NO=Number of Ontologies

Topic	Example Terms	F	NO
State	busy, resume, standby, waiting, closing, pause, idle, stop, sleep, timeout	482	16
System	system, service, architectural, software, hardware, middleware, application	308	15
Home safety & security	intrusion, siren, warning, alarms, alerts, notification	291	5
Parameter	parameter, input, output, variable, return, type, set, index, message	288	21
Measurement	measure, observation, sensing, actuation, acquisition, test	279	15
Engineering parameters	capacity, reliability, availability, load, storage, efficiency, occupied, interrupted	278	17
Appliance	smartphone, server, computer, device, sensor, actuator, platform, pc	198	19
Actuator	board, command, adapter, Modbus, clock, playback, button, display, gate	195	5
Ambient parameters	sound, luminosity, noise, radiation, humidity, temperature, co2, motion, smoke, sound	193	15
Climate	weather, climate, forecast, cloudy, rainy, sun, dry, cold, hurricane, rain, river, flood	192	10
Event	action, task, activity, agent, interaction, agent, entity	191	16
Metadata	biometric, information, history, update, read, record, picture, track, status, report, data	183	16
Health	blood, stress, pulse, heart, glucose, pressure, cholesterol, heartrate, calorie, ph, skin	150	19
Network	zigbee, insteon, bridge, dimmer, bticino, zwave, konnex, Bluetooth, gateway, rtu, plc	146	14
Identifiers	ip, mask, port, username, password, uuid, identification, label, id, name, tag	145	17

TABLE III: The most important weather terms among 24 different ontologies.

Topic	Example Terms	F	NO
Water related weather phenomena	evapotranspiration, pillow, wetting, falling, melt, lwe	139	10
Weather observation & forecast	radar, station, meteorological, measurement, observation, weather	129	10
Physical quantities & mechansims	entropy, gradient, diffusion, acceleration, viscosity, adiabatic	110	10
Weather forecasting sensors	solarimeter, thermocouple, cylindrical, albedometer, pyrgeometer	102	2
Sun-influenced states and phenomena	dry, irradiation, shade, wet, wetness, hot	100	10
Weather status	fair, snowy, rainy, wet, showers, sleet	82	8
Measurement	quantity, relative, variance, error, power, deviation	80	11
Weather reports	status, reported, reports, merged, service, event	78	8
Physical thing	thing, entity, physical, phenomena, property, description	72	13
Geospatial	elevation, angle, near, deviation, land, longitude	62	8
Sun-influenced states and phenomena	moisture, heat, vapor, volume, energy, liquid	62	8
Soil	soil, atmospheric, flux, source, surface, combined	48	8
Time	instant, pulse, hour, duration, intensity, timed	47	7
Atmospheric circulation	atmosphere, convection, circulation, condensation, layer, convective	47	10
Metadata	provider, status, qualified, dated, rights, comment	45	9

TABLE IV: The most important smart city terms among 24 different ontologies.

Topic	Example Terms	F	NO
Country statistical profile	revenues, households, sales, gdp, expenditures, population, exports	216	1
Smart city topics	mobility, tourism, grid, urban, renewable, city, technological, innovation, ict	100	3
Dimension	size, scale, volume, depth, load, diameter, weight, width, length, density, age, area	82	8
Transport & logistic	logistics, supply, financial, transportation, transport, enterprise, market, government	70	6
Data collection	sensor, data, stream, observation, measurement, metadata, knowledge	69	5
Context	neighbourhood, outdoor, indoor, buildings, retail, domestic, land, district, commercial, industrial	67	2
Pipes	pipe, material, joint, coating, lining, rubber, mass, products	67	2
IoT	object, internet, thing, thing, entity, device, service, physical, virtual, cloud, platform, web	62	7
Specification	components, definition, term, feature, concept, element, function, aspect, class, type, property	59	7
Services	system, operations, applications, infrastructure, equipment, sys, services	52	4
Metadata	id, info, value, tag, timestamp, belongs, producer, uri,	52	7
Transportation mode	vehicles, taxi, bike, motorcycle, car, bicycle, commuters, parking	50	5
Public utilities & buildings	water, food, waste, solid, disposed, hazardous	49	3
Buildings & infrastructure	churches, offices, clinics, nursing, sport, garden, parks, entertainment, buildings, retail, office	49	4
Economy	workforce, jobs, trade, monetary, economy, investement, employment, labour, productivity	48	2

among the list of available smart city ontologies. It is also interesting to see that clusters like *Building & Infrastructure* or *Transportation Mode* are essential terms in the smart city domain.

B. Discussions

The smart home results (Section IV-A1) generate clusters such as *State (state, busy, waiting, closing, stopping, pause, stop, etc.)* which demonstrates the need to understand the state of the devices. The term frequency values in the smart home domain are higher than the other domains. It shows that it is actively researched and has a higher industry interest compared to other IoT application domains. The weather results (Section IV-A2) demonstrate interesting clusters such as *Time (instant, pulse, hour, duration, intensity, timed, frequency, interval, time, series)* which demonstrates the need for temporal aspects. The cluster *Weather (fair, snowy, rainy, wet, showers, etc.)* illustrates the possible weather states. The smart cities results (Section IV-A3) show the cross-domain disciplinary since it involves transportation. For instance, the cluster *Transportation Mode (travel, motorized, vehicles, taxi, bike, etc.)* has been found.

TABLE V: List of ontologies selected for evaluation per domain.

Application Domain	Ontologies
Smart Home	dogont [23], Dem@Care [24]
	Codamos [25], activity recognition [26]
Smart City	Airport ontology [27], vital EU FP7 [28]
	CityPulse EU FP7 Stream Annotation ontology (SAO) [29]
Weather	Staroch [15], ThinkHome [30], SemSOS [31]

V. Evaluation

The objective of this evaluation is to identify if the most frequent topics provided by the KE4WoT methodology per application domain (home, weather, city) can sufficiently describe the main content of existing ontologies. The proposed methodology is evaluated in an empirical study with 23 domain experts which include an analysis that gives a complete overview of the performance of the descriptiveness of the most frequent topic.

A. Ontology Selection

10 ontologies representative of 3 different domains in the IoT field are collected from LOV4IoT for evaluation purposes (Table VI). The most important ontologies in each domain are selected according to those criteria:

- Citations of the scientific publications describing the ontology (e.g., the SSN ontology V1 [11] has more than 1000 citations). Higher is the number; better the ontology might be. However, this criterion cannot be applied to recent publications.
- Journal impact factor and conference ranking. Higher the ranking is, better would be the ontologies.
- Recent publications increase the chance to have the authors maintaining ontology and integrating previous ontologies.
- Ontologies disseminated in standardizations (e.g., W3C Web of Things ontology[14], W3C SSN/SOSA ontology [21], ETSI M2M SAREF ontology [22]).
- The ontology-based projects are considered more impactful when industrial partners are involved: the implementation is more reliable.
- Domain experts involved (not computer scientists) sharing their expertise.
- Ontology code (e.g., downloadable). In science, the experiments should be replicable[15].

B. Materials

23 domain experts participated in the questionnaire available online[16] to receive a ground truth. Experts are either directly involved in developing IoT ontologies or were an open audience having the domain expertise to describe each ontology using three keywords. We checked the participant expertise level in our three selected IoT application domains (i.e., weather, city, and home) and knowledge engineering. Experts rated their experience in a Likert scale of five levels, from 'totally disagree' to 'totally agree.' Most of the experts had a strong knowledge in the field of IoT and knowledge engineering, but an average experience in weather and city domains.

The experts analyze the list of ontologies (through a series of figures from the ontology classes in Protege) in our selected domains to select the top three keywords that best describes that ontology in relation to the keywords that were obtained from the clusters given by KE4WoT. The total number of keywords which was available for the domain expert to choose from were: 28 for smart city, 21 for weather, and 26 for smart home ontologies. The questionnaire provided the main topics and the list of topic members to give the meaning of each keyword to reduce ambiguity.

C. Results

The evaluation investigates whether the topics identified by KE4WoT are sufficiently descriptive to be applied to ontologies. We employ sets of keywords manually assigned by human domain experts as ground truth. We combine the most important topics identified by KE4WOT and the list of ontologies, mapping each individual concept of the ontologies to its corresponding cluster to automatically identify the top three keywords and use these to describe the ontology. The ground truth is aggregated from survey responses of domain experts who used the KE4WoT topics to assign three keywords per ontology using majority consensus: for each ontology to be described, the classification $C \subset K \times F$ are tuples (k_i, f_i) where $k_i \in K$ is a keyword which the domain experts selected and f_i the relative frequency of the keyword among domain experts. The three keywords for the ground truth are then identified applying majority consensus as in 1 repeatedly and removing (k^*, f_i) from C until the three keywords with the highest relative frequencies are found.

$$k^* = \underset{k,(k,f_i) \in C}{\arg\max}\, f_i \tag{1}$$

Table VI shows the percentage of similarity per ontology between the set of topics identified using KE4WoT and the ground truth. Keywords are highlighted in bold when there is an agreement between groundtruth and KE4WoT. The results shows that KE4WoT was able to identify the majority of the keywords in the ground truth with a total percentage of 66%. From the 21 (weather) to 28 (smart city) available keywords, in 7 of the 10 ontologies at least two of the three assigned keywords match, all 10 ontologies show at least one match. For Staroch ontology, the toolchain was able to identify three matching keywords.

The percentage of similarity in the last three ontologies may seem low at first glance. To further investigate the lower similarity levels in the CODAMOS, konlarkon and airport ontology, we calculated the inter-rater agreement (degree of agreement or disagreement) between the domain experts opinion using Entropy E measure as in Equation 2:

[14] https://www.w3.org/TR/wot-thing-description/

[15] https://goo.gl/dmR4hB

[16] https://bildungsportal.sachsen.de/survey/limesurvey/index.php/716626/lang-en

TABLE VI: Percentage of similarity between KE4WoT and ground truth

Ontology	KE4WoT & ground truth	% of similarity
staroch	**weather observation & forecast** **sun-influenced states & phenomena** **weather report**	100
dogont	**state**, home safety & security, **system**	66.6
Dem@Care	state, **time**, **activities**	66.6
kofler	weather observation & forecast **weather report** **sun-influenced states & phenomena**	66.6
semsos	**weather observation & forecast**	66.6
	physical thing, **measurement**	66.6
VITAL	**IoT**, **data collection**, operations	66.6
citypulse	mathematic, **data collection**, **sensing**	66.6
CODAMOS	**system**, geospatial, engineering parameters	33.33
konlarkon	identifier, event, **appliance**	33.33
airport	**cargo**, metadata, emergency operations	33.33

$$E = -\sum_{i=1}^{n} f_i \lg f_i \qquad (2)$$

and the normalized Herfindahl dispersion measure H* as in Equation 3:

$$H* = \frac{n}{n-1}(1 - \sum_{i=1}^{n} {f_i}^2) \qquad (3)$$

based on the relative frequencies f_i of the keywords assigned by human experts for $n = 17$ keywords used. Entropy and Herfindahl measures indicate the disorder or dispersion between domain experts keyword selection. The higher the disagreement between the domain experts, the more different keywords selected, the closer E and H* get to 1. The average Entropy was calculated at 0.7 and the average Herfindahl measure at 0.9 for the last three ontologies. It clearly shows the low consensus and different opinions between the domain experts in these ontologies.

The evaluation results are encouraging. They show that the identified main topics of IoT ontologies from the three different domains can be used as keywords to describe sufficiently existing ontologies.

VI. Related Work

The statistical analysis of IoT ontologies and application domains is missing. The SSN ontology validator[17], focused on the usage of the W3C SSN V1 ontology. It validates 7 SSN-based ontologies and datasets and generates a term tag cloud, showing the recurrence of the terms, based on the ontology analysis.

Semantic Interoperability for IoT issues are highlighted in [7] [40]. The **European Research Cluster on the Internet of Things AC4** released IoT semantic interoperability best practices and recommendations [6], but does not reference concrete tools to encourage the reuse of the domain knowledge already designed. Ontology matching tools [41] have been considered with IoT ontologies. However, due to the lack of best practices (e.g., no rdfs:label or rdfs:comment) ontology alignments are not successful.

Knowledge Extraction Processes: There is no research yet doing term analytics on IoT ontologies using the word2vec approach. Bizer et al. [42] analyzed the usage of RDFa, Microdata, and Microformats vocabularies within HTML web pages. A similar analysis is missing to understand semantic-based IoT annotations. LODStats statistics[18] and SPORTAL statistics[19] compute top-k classes and properties (e.g, `rdfs:type` and `rdfs:label` are the most employed properties) but do not provide any information about the domain itself.

Knowledge Repositories: We designed the LOV4IoT [43] ontology catalog to statistically analyze IoT term analysis since LOV [44], BioPortal [45], READY4SmartCities [46], have limitations highlighted in [17]. Scooner [47] analyzed abstracts from Pubmed and aligned the main keywords with health datasets from the Linked Open Data Cloud but is limited since it does not cover scientific libraries for IoT. KBpedia[20] aggregates structured datasets (Wikipedia, Wikidata, schema.org, DBpedia, GeoNames, OpenCyc, UMBEL) but is limited for IoT, smart cities, smart homes, and weather domains. Tirado et al. [48] highlight that the lack of standard data models and structures forces developers to create models from scratch. Developers need collaborations with domain experts having the correct background knowledge.

Information Retrieval: There are some solutions in the literature for finding and ranking ontologies to help users select the most appropriate ontology for reuse (as illustrated in Table VII). However, most of these solutions are no longer supported or do not reference IoT ontologies. Our work differentiates itself with these solutions by extracting the individual ontology resources and grouping the similar resources to identify the different topics covered in each ontology.

The limitations of the existing literature are:

- Lack of IoT statistical analysis to detect common knowledge and recurrent IoT terms.
- Lack of common agreement on IoT ontologies for building iot.schema.org and extend the W3C SSN/SOSA ontology.
- Table VII references ontology selection and ranking tools, but they are not mature enough for identifying IoT related ontologies.
- Developers need domain experts to build ontologies. Automatic tools analyzing existing ontologies in the same topic could partially replace domain experts in the development phase. Domain experts are still required for the ontology evaluation phase.

VII. Conclusion and Future Work

Our KE4WoT methodology studied IoT ontologies using word2vec and k-means machine learning techniques, to statistically analyze the most common topics (concepts and properties). The methodology is applied to three IoT application domains: smart home, smart city, and weather. The identified topics have been discussed within the iot.schema.org meeting which allows our findings to be used as a starting point for

[17] http://iot.ee.surrey.ac.uk/SSNValidation/about.html

[18] http://lodstats.aksw.org/stats

[19] http://www.sportalproject.org/Sadgraphs.html

[20] http://kbpedia.org/

TABLE VII: Qualitative comparison between existing information retrieval solutions. Legend: K=Keyword, C=Class, SC=Subclass, L=Label, P=Property, D/R= Domain/Range, Li=Literal, CM=Comment, CO=Corpus, RO=Ranked Ontologies

Approach	Input	Ranking criteria	Search terms	Tool Availability	IoT Support	Output
OntoKhoj [32], **Sindice [33]**	K	semantic links, coverage	C, SC, D/R	×	×	RO
Swoogle [34], Watson [35]	K	semantic links, coverage	C, P, L, CM, Li	✓	×	RO
AKTiveRANK [36]	K	ontology structure	terms	×	×	RO
Falcons [37]	K	concept popularity	C, P, L	×	×	RO
OntoSelect [38], **OntoSearch2 [39]**	K, C	connectness, ontology structure	CO, P, L	×	×	RO
KE4WOT	ontologies	concept popularity	C, SC, P, L, D/R	✓	✓	popular concepts

further investigation towards generating an integrating knowledge graph. KE4WoT helps researchers to stay updated with the scientific literature and the content of existing ontologies.

As future work the KE4WoT methodology can be applied to additional domains such as transportation, robotic, IoT-based healthcare, affective science, disaster management, etc. The popular topics, extracted from ontology datasets, would help to automatically generate an integrated schema which would be relevant for standardizations (W3C WoT, OneM2M, etc.) or iot.schema.org. Ontology restrictions could be analyzed to design IF THEN ELSE rules to interpret IoT data.

References

[1] Atzori, L., et al.: The internet of things: A survey. Computer networks (2010)

[2] Guinard, D., Trifa, V.: Building the web of things: with examples in node. js and raspberry pi. Manning Publications Co. (2016)

[3] Sundmaeker, H., Guillemin, P., et al.: Vision and challenges for realising the internet of things. Cluster of European Research Projects on the Internet of Things, European Commission (2010)

[4] Noura, M., Atiquzzaman, M., Gaedke, M.: Interoperability in Internet of Things Infrastructure: Classification, Challenges, and Future Work. (2017)

[5] Noura, M., Atiquzzaman, M., Gaedke, M.: Interoperability in Internet of Things: Taxonomies and Open Challenges. Mobile Networks and Applications (2018)

[6] Serrano, Martin and Barnaghi, Payam et al.: Internet of Things IoT Semantic Interoperability: Research Challenges, Best Practices, Recommendations and Next Steps. Technical report, European Research Cluster on the Internet of Things, AC4 (2015)

[7] Murdock, P., et al.: Semantic Interoperability for the Web of Things (White Paper) (2016)

[8] Barnaghi, P., Wang, W., Henson, C., Taylor, K.: Semantics for the Internet of Things: Early Progress and Back to the Future. International Journal on Semantic Web and Information Systems (IJSWIS) (2012)

[9] Noura, M., Gyrard, A., Heil, S., Gaedke, M.: Concept Extraction from the Web of Things Knowledge Bases. In: International Conference WWW/Internet 2018, Elsevier (2018) Outstanding Paper Award.

[10] Kosala, R., Blockeel, H.: Web mining research: A survey. ACM Sigkdd Explorations Newsletter **2**(1) (2000) 1–15

[11] Compton, M., Barnaghi, et al.: The ssn ontology of the w3c semantic sensor network incubator group. Web Semantics: Science, Services and Agents on the World Wide Web (2012)

[12] Gyrard, Amelie and others: Standardizing generic cross-domain applications in Internet of Things. In: IEEE Globecom Workshop on Telecommunications Standards, From Research to Standards. (2014)

[13] Agarwal, R., Fernandez, D.G., Elsaleh, T., Gyrard, A., et al.: Unified IoT ontology to enable interoperability and federation of testbeds. In: IEEE WF-IoT. (2016)

[14] Ruta, M., et al.: Machine Learning in the Internet of Things: a Semantic-enhanced Approach. Semantic Web Journal (2017)

[15] Staroch, P.: A weather ontology for predictive control in smart homes. Master's thesis (2013)

[16] Gyrard, A., et al.: Reusing and Unifying Background Knowledge for Internet of Things with LOV4IoT. In: IEEE FiCloud. (2016)

[17] Gyrard, A., et al.: Building IoT based applications for Smart Cities: How can ontology catalogs help? IEEE IoT Journal (2018)

[18] Kitchenham, B.: Procedure for undertaking systematic reviews. Computer Science Depart-ment, Keele University and National ICT Australia Ltd, Joint Technical Report (2004)

[19] Lécué, F., other: Semantic traffic diagnosis with star-city: Architecture and lessons learned from deployment in dublin, bologna, miami and rio. In: ISWC 2014. Springer (2014)

[20] Komninos, N., et al.: Smart city ontologies: Improving the effectiveness of smart city applications. Journal of Smart Cities (2016)

[21] Haller, A., Janowicz, K., Cox, S., Le Phuoc, D., Taylor, K., Lefrancois, M.: Semantic Sensor Network Ontology. W3C Recommendation (2017)

[22] Daniele, L., Solanki, M., den Hartog, F., Roes, J.: Interoperability for smart appliances in the iot world. In: ISWC, Springer (2016)

[23] Bonino, D., Corno, F.: Dogont-ontology modeling for intelligent domotic environments, Springer (2008)

[24] Ye, J., et al.: Semantic web technologies in pervasive computing: A survey and research roadmap. Pervasive and Mobile Computing (2015)

[25] Preuveneers, D., et al.: Towards an extensible context ontology for ambient intelligence. In: Ambient intelligence. Springer (2004)

[26] Wongpatikaseree, K., et al.: Activity recognition using context-aware infrastructure ontology in smart home domain. In: Knowledge, Information and Creativity Support Systems Conference, IEEE (2012)

[27] Park, K., et al.: Semantic reasoning with contextual ontologies on sensor cloud environment. Journal of Distributed Sensor Networks (2014)

[28] Kazmi, A., Jan, Z., Zappa, A., Serrano, M.: Overcoming the heterogeneity in the internet of things for smart cities. In: International Workshop on Interoperability and Open-Source Solutions, Springer (2016)

[29] Kolozali, S., et al.: A knowledge-based approach for real-time iot data stream annotation and processing. In: IEEE iThings conference. (2014)

[30] Kofler, M., et al.: A semantic representation of energy-related information in future smart homes. Energy and Buildings (2012)

[31] Henson, C., et al.: Semsos: Semantic sensor observation service. In: Collaborative Technologies and Systems Symposium, IEEE (2009)

[32] Patel, C., et al.: Ontokhoj: a semantic web portal for ontology searching, ranking and classification. In: Web Information and Data Management Workshop, ACM (2003)

[33] Tummarello, G., Delbru, R., Oren, E.: Sindice.com: Weaving the Open Linked Data. (2007)

[34] Ding, L., Finin, T., Joshi, A., et al.: Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the International Conference on Information and Knowledge Management, ACM (2004)

[35] d'Aquin, Mathieu and Motta, Enrico: Watson, more than a semantic web search engine. Semantic Web (2011)

[36] Alani, H., Brewster, C., Shadbolt, N.: Ranking ontologies with AKTiveRank. In: International Semantic Web Conference, Springer (2006)

[37] Cheng, G., et al.: Falcons: searching and browsing entities on the semantic web. In: WWW, ACM (2008)

[38] Buitelaar, P., et al.: OntoSelect: A dynamic ontology library with support for ontology selection. In: ISWC Demo, Citeseer (2004)

[39] Pan, J.Z., Thomas, E., Sleeman, D.: Ontosearch2: Searching and querying web ontologies. Proc. of WWW/Internet (2006)

[40] Gyrard, A., et al.: A survey and analysis of ontology-based software tools for semantic interoperability in IoT and WoT landscapes. In: IEEE World Forum on Internet of Things (WF-IoT). (2018)

[41] Euzenat, J., Shvaiko, P.: Ontology matching 2nd Edition. Springer-Verlag (2013)

[42] Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of rdfa, microdata, and microformats on the web–a quantitative analysis. In: ISWC, Springer (2013)

[43] Gyrard, A., et al.: LOV4IoT: A second life for ontology-based domain knowledge to build Semantic Web of Things applications. In: IEEE FiCloud. (2016)

[44] Vandenbussche, P.Y., et al.: Linked Open Vocabularies (LOV): a Gateway to Reusable Semantic Vocabularies on the Web. Semantic Web Journal (2016)

[45] Whetzel, P., Noy, N.e.a.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic acids research (2011)

[46] Garcia-Castro, R., et al.: Ready4smartcities: Ict roadmap and data interoperability for energy systems in smart cities. In: ESWC. (2014)

[47] Kavuluru, R., et al.: An up-to-date knowledge-based literature search and exploration framework for focused bioscience domains. In: Health Informatics Symposium, ACM (2012)

[48] Tirado, J.M., Serban, O., Guo, Q., Yoneki, E.: Web data knowledge extraction. Technical Report - University of Cambridge (2016)