Topical Search Engine for Internet of Things

Mosab Faqeeh, Mahmoud Al-Ayyoub, Mohammad Wardat, Ismail Hmeidi and Yaser Jararweh
Jordan University of Science and Technology
Irbid, Jordan
Emails: {mos3b.faqeeh, malayyoub, mawardat12}@gmail.com, {hmeidi, yijararweh}@just.edu.jo

***Abstract*—Internet of Things (IoT) has become a common buzzword nowadays in the Web. However, there is no search tool currently in place for discovering and learning about the different types of IoT elements. Hence, this paper presents a topical search engine for IoT. The motivation for a topical search engine comes from the relatively poor performance of general-purpose search engines, which depend on the results of generic Web crawlers. The topical search engine is a system that learns the specialization from examples, and then explores the Web, guided by a relevance and popularity rating mechanism. The results show that the proposed topical search engine outperforms other general search engines.**

I. Introduction

The Internet of Things (IoT) is novel paradigm equipping everyday objects (food packages, garments, furniture, paper documents, plantations, livestock, etc.) with adequate technology to generate new data and communicate with other objects in order to reach common goals [1]. The authors of [2] spawned the term of "Internet of Things" from the idea of Radio Frequency Identification Tags (RFID), which still stands at the forefront of the technologies driving the vision. Syntactically, the expression is composed of two concepts: "Internet" and "Thing". The former implies a network oriented vision of IoT, while the latter implies an object that is not precisely identifiable. In fact, IoT semantically means "a world-wide network of interconnected objects uniquely addressable, based on standard communication protocols" [3].

IoT is expected to touch different aspects of the everyday life and behavior of potential users. It will become one of the most vital topics in several industrial and academic sectors. Most effects of IoT introduction will be visible in both working and domestic fields. In fact, many programs and applications are moving towards IoT as it is considered as part of the overall Internet of the future. They are readily available and easy to access, even in somewhat remote regions [4].

IoT represents the next evolution of the Internet that will play a leading role in the education, communication, business, science, government, and humanity [5]. Therefore, it is expected that there will be numerous Web pages about IoT in the near future with a large number of requests for such pages. Such rise of requests may produce a problem if any user wants to determine the most suitable Web pages among the many ones available. This is especially important due to the number of items involved in the future Internet is destined to become extremely high. Therefore, issues related to how to represent, store, interconnect, search, and organize information related to IoT will become very challenging. However, there are no principles, protocols, or search mechanisms for discovering different types of pages that support IoT technologies, and users are forced to look through those pages manually. Hence, there is a need for a search engine customized for IoT.

Search engines are one of the main mechanisms by which users obtain information on the Web. For any individual wanting information about something on the Internet, search engines are the first choice to help find what the user wants. In our case, we focus more on specialized search engines, which work based on indexed pages for particular topics only. These search engines feature superior speed and accuracy. There are often many topics for which general-purpose search engines, such as Google and Yahoo, might return irrelevant results. This is due to the confusion created by the similarity in the terminology of different unrelated topics.

The reason we use a specialized search engine instead of an existing, general purpose one, such as Google, is that trying to use a general search engine to find a specific pages of IoT technologies would yield a large number of irrelevant results, making it nearly impossible to sift through all the unrelated ones to find what you want. Here, there is a need to use a specialized search engine to give a high accuracy and speed in the extraction of relevant results. Therefore, this paper aims to design and build a specialized search engine for IoT technologies keywords available on the Web, and presents the results to the user. In addition, it is worth noting that we found no previous attempts to develop a specialized search engine on IoT technologies.

The proposed search engine is written as a Web directory for IoT technologies, i.e. giving the users the opportunity to attach their own Uniform Resource Locators (URLs), into the system. Moreover, it is equipped with a special classification tool to help improve the search results. As shown in Section IV, the proposed search engine already shows impressive results. It is currently being improved in terms of both speed and accuracy in addition to the capability of serving many concurrent users by holding concurrent sessions. Since database connectivity is a big bottleneck, a method which reduces database workload as much as possible is being implement.

The rest of this paper is organized as follows. The following section gives a general overview of the current literature on IoT and topical search engines. The proposed system is discussed in Section III and tested in Section IV. Finally, concluding remarks along with a discussion of future work is discussed in Section V.

978-1-4799-7100-8/14/$31.00 ©2014 IEEE

II. Related Works

A. Internet of Things (IoT)

There are broad areas of IoT in both industrial and development of applications to enhance and improve the quality of our lives: at home, at work, when sick, while travelling, when jogging and at the gym. There are many beneficial applications provided by IoT technologies that can be grouped mostly into the following domains [4]:

- Transportation and logistics domain.
- Health care domain.
- Smart environment (home, office, plant) domain.
- Personal and social domain.

The current Internet that we use today is dramatically different from the Internet of the future. Tt is a purely transparent carrier of packets mostly used for the publishing and retrieving information. Therefore, several users use middle boxes to improve security and accelerate applications leading to the concept of Data-Oriented Network Architecture (DONA) proposed in [6]. According to such a concept, data and the related queries are self-addressable and self-routable.

According to [7], IoT would include 50 to 100 trillion objects and it should be able to follow their movement and manage their generated data. The estimated number of Web page that will be available as part of IoT is around 17 billion Web pages. Therefore, several new problems related to IoT technologies need to be addressed. One such problem is the search problem which is critical for the next Internet generation. The challenges of this problem include how to represent, store, interconnect, search, and organize information generated by IoT objects. In this context, semantic technologies could play a key role in the IoT visions [8], [9], [10], [11].

Several works envisioned concepts coinciding with IoT. For example, in what they call "Web of Things" vision, the authors of [12] proposed an approach to connect and integrate the services offered by devices and objects in the real world with the Web allowing them to be flexibly combined with other virtual and physical resources. In another insightful work, the authors of [13] discussed Web Squared, which is an evolution of the Web 2.0 the same author discussed in 2003. It is aimed at integrating Web and sensing technologies together . Our phones and cameras are being turned into eyes and ears for applications; motion and location sensors (microphone, cameras, GPS, etc.) tell us where we are, what we are looking at, and how fast we are moving. Data is being collected, presented, and acted upon in real time. The collective intelligence applications being developed under IoT depend on managing, understanding, and responding to such massive amounts of data.

One early examples of such application is mobile ticketing. In [14], the authors investigated mobile interaction with tagged, everyday objects and associated information (description, costs, and schedule) that is based on IoT and its technologies. Two prototypes for mobile interaction with smart posters build to realize multi-tag interaction with physical user interfaces. The proposed framework allowed the user to get information about several categories of options from the Web by either pointing the mobile phone to the visual markers, or hovering it over the NFC tag. Then the mobile phone gets the relevant information (stations, numbers of passengers, costs, available seats and type of services) from Web services, and allows the user to buy the relevant ticket.

In the monitoring environmental parameters domain, several works [15], [16] suggested studying the quality of perishable goods (such as fruits, meat, dairy products and fresh-cut produce) under different issuing policies. Such issues are vital parts of our nutrition. Sensor technologies and pervasive computing can significantly improve the efficiency of the food supply chain.

In the health care domain, the authors of [17] stated that there are four functional domains can be supplemented or complemented by the IoT technologies: tracking (of objects and people such as staff and patients), identification and authentication (of patients, infants, etc.), automatic data collection and sensing.

According to the previously discussed studies, IoT technologies are destined to grow and flourish as they have been recognized as the future of the Internet. Hence, with the increased interest and the increased connectivity, more and more data need to be managed Accordingly, there is a need to create specialized search engines that periodically insert information to be used for IoT purposes.

B. Search Engines

Due to their significance, search engines have received a considerable amount of interest over the past three decades. The focus of this paper is on one specific type of search engine, which is specialized (topical) search engines. Thus, our discussion here will be limited to such search engines.

Information retrieval in medical fields currently constitutes a large portion of Web searches. Currently, much of the general population, as well as experts are looking for health care and medical material at their homes without requiring much effort, through the Internet. Medical Information Retrieval System (MIRS) over search engines gives high quality information to the user using fixed questionnaires systems. In [18], the authors built a model for inexperienced users with small amount of knowledge to view the system by choosing from listed related results. Along with the framework they implemented, they created an Intelligent Medical Search Engine (IMSE) for examining health material on the Web. The operation system of IMSE contained a health ontology and questionnaire to enable inexperienced users to use the system via the Internet. IMSE presented and covered professional system tools interested in the search engine field. It contained several key methods to increase its performance and search outcome quality.

Another type of search engines is semantic search engines.One of the better known of these search engines is Hakia [19], a common determination semantic search engine that examined structured text such as Wikipedia. Hakia itself is a "meaning-based (semantic) search engine". The authors focused on delivering search outcomes using meaning similarity,

more than by the similarity of search keywords themselves. The offered news, blogs, and social network articles, are treated by Hakia's branded central semantic methodology titled QDEXing [20]. It is able to treat different kinds of alphanumeric articles by their Semantic technology with help from third party API feeds [21].

Focused crawler, originally presented by de Bra et al. [22] and then studied by several others [23], [24], are intended to fetch results appropriate to a pre-specied area of concern by the Web's hyperlink construction. A focused crawler begins with a seed's gradient of relevant URLs. URL keywords and page content relevance are naturally explored in evaluation link rate. McCallum et al. [25] build a model using Naive Bayes classiers to classify hyperlinks, and Diligenti et al. [26] build the context and body diagram knowledge to direct a focused crawler. More than exploratory related nodes at each time, these methods qualified an apprentice with structures collected from source chief active to the related nodes. Chakrabarti et al. [24], instead, offer another methodology in hypertext diagrams counting in-neighbors (documents mentioning the objective document) and out-neighbors (documents that objective document mentions) as key to many classiers. From these authors, a focused crawler is able to obtain related pages gradually, though a typical crawler rapidly indexes a huge amount of unrelated pages and misplaces its method.

III. Methodology

Search engines start by crawling Web pages gather information to be stored in a specific storage space for additional processing. A crawler is a program for which we specify a seed URL, and keep going based on that URL retrieving connected pages.[1] A page is parsed for additional URLs by URL Normalization, where these URLs are saved in storage for crawling [27], [28], and are used to retrieve the more available Web pages from a Web server. The process of crawling may be divided among multiple distributed crawlers.

Once a document is indexed, it is capable of being searched for using an identifying query. The interface then takes the input from the user and determines the best result for it. The user query may be passed through additional optional modules, such as spell correction.

Once the query is processed based on the indexed information, the outcomes retrieved and ranked based on relevance to the query. The relevance of the search results is then determined; one such metric may be the search term frequency [29]. A search engine may take into consideration the search term frequency in combination with its position in the site document, as well as its presence in the Web site information or metadata (words that appear near the beginning of the document may be given extra weight)

Search engines can either be general purpose or specialized (topical search engines) depending on whether they are specialized in searching for information related to one specific or not. As discussed in Section II, several topical search engines have been proposed for different fields. This paper aims to design a topical search engine for all IoT related Web sites accessible on the Internet. It has the potential to attach one's own links to the system. It performs searches by keywords as well as text and it is equipped with a special classification tool to help improve the search results. Also, the whole system is optimized for speed and accuracy.

This section describes the details of the underlying structure and how every part is implemented. A committed server is dedicated for this project. The server runs on an IBM cloud platform operated by the Jordan University of Science and Technology (JUST). Since we are using the concept of virtualization, described in detail in [30], a moderately fast processor and a lot of memory was required. We choose a 1.4 GHz AMD processor with 1.5 GB of main memory, which should be enough. The server has a 10/100 Mbit network card that is connected to a 2.0 Mbit Internet connection. The server works with two 40 GB hard drives; one which holds data and another for backup purposes. The server's operating system is Windows 7 Ultimate. The Web server, XAMPP, is free of charge and open source Web server key stack package. The database is MySQL 2010, which is an extremely fast database with a pleasant interface. The programming language of choice for building the crawler and the search engine is PHP. It allows for the construction of active Web pages. The tables in database are built extremely efficiently; the crawlers, the search engine, and the indexing algorithms, all divide the collective memory space. Also, since PHP is an active programming language, it is probable to construct connections that keep in touch straight to the SQL tables. The system may also be ported to a diverse set of operating systems easily, because both PHP and XAMPP are flexible and portable.

Architecture The architecture of the proposed search engine is shown in Figure 1. As typical with search engines, the first step is crawling the Web looking for Web pages whose contents are related to IoT. This is achieved using several distributed crawlers. The pages they collect are parsed to determine keywords they contain. The URLs of these pages are indexed and stored in a database along with their keywords. When a new query comes in from a user, the search engines needs to determine which keywords exist in the query and match these keywords with the ones stored in the database. The matched keywords are then analyzed to determine the score of a certain page. The pages are then displayed to the user ranked according to their scores.

A search engine needs to be user-friendly. Our proposed search engine provides a simple an efficient user interface. Moreover, it helps the user writing his/her query by providing him/her with a list of suggested keywords as extracted from the collected pages as shown in Figure 1.

One of the most important components of a search engine is the ranking and indexing module. This module is responsible for parsing Web pages, matching keywords and counting the number of hits (keyword count) or the regularity of the keywords in a specific page. These computed values are used

[1] http://en.wikipedia.org/wiki/WebCrawler

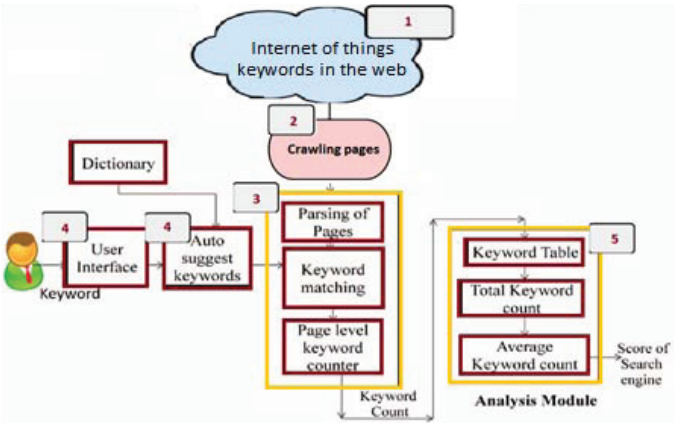

Fig. 1. High level system architecture of the prototype search engine.

in the analysis part to determine the score/rank of a certain page.

Ranking Technique. There are many ranking techniques proposed in the literature. The choice of the best ranking techniques depends on many factors such as dataset, performance, simplicity, and computational effectiveness. In this work, we use Term FrequencyInverse Document Frequency (tf-idf) techniques, which is an mathematical guide proposed to return how essential a word is to a document in a corpus[31]. It is frequently used as a weighting feature in information retrieval and text mining. The tf-idf rate increases in proportion to the amount of times a word appears in the document. However, the rate is adjusted by taking the occurrence of the word in the corpus into consideration to account for the fact that some words are more common than others.

As its name suggests, tfidf is a combination of two statistics, term frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist. In the case of the term frequency $\mathrm{tf}(t,d)$, the simplest choice is to use the raw frequency of a term in a document, i.e. the number of times that term t occurs in document d. If we denote the raw frequency of t by $\mathrm{f}(t,d)$, then the simple tf scheme is $\mathrm{tf}(t,d) = \mathrm{f}(t,d)$. Other options include boolean frequencies in which $\mathrm{tf}(t,d) = 1$ if t occurs in d and 0 otherwise. Also, logarithmically scaled frequency can be used in which $\mathrm{tf}(t,d) = \log(\mathrm{f}(t,d)+1)$. A third option is to use augmented frequency to prevent a bias towards longer documents, e.g. raw frequency divided by the maximum raw frequency of any term in the document:

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

The inverse document frequency is used to determine how much information the word represents, i.e., whether the term is general or rare throughout all documents. It is the logarithmic measure of the fraction of the documents that include the word, obtained by dividing the total amount of documents by the amount of documents containing the term, and then calculating the logarithm of that quotient.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Mathematically the base of the log function does not matter and constitutes a constant multiplicative factor towards the overall result. Then tfidf is calculated as

$$\text{tf-idf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

A high weight in tfidf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Since the ratio inside the idf's log function is always greater than or equal to 1, the value of idf (and tf-idf) is greater than or equal to 0. As a term appears in more documents, the ratio inside the logarithm approaches 1, bringing the idf and tf-idf closer to 0.

We tested our tf-idf implementation on a collection of 210,000 documents (pages) from our database. These pages are gathered by our crawler along with their URLs. After we apply this technique, URLs would have ranks and vectors referring to their content. When the user types a query, the proposed search engine computes the tf-idf of this query and formulates its vector which is compared with the existing vectors. Table I shows a sample outcome from the tf-idf technique.

TABLE I
SAMPLE TABLE OF TF-IDF OVER SOME OF COLLECTIONS.

Doc Collection	No. Docs	$\sum \mathrm{tf}(t,d)$	Avg. of $\mathrm{tf\text{-}idf}(t,d,D)$
1	50,000	4,000	0.8
2	25,000	2,500	0.6
3	80,000	6,000	0.9
4	10,000	1,200	0.5
5	10,000	1,100	0.5

IV. Results and Evaluation

The collected database currently contains about 210,000 URLs. This number can be easily increased by running the crawlers for longer periods of time.We test our search engine by submitting various queries related to IoT and manually evaluate the returned pages (i.e., the results of the search) to determine which ones are relevant (hits) and which ones are not (misses). The search queries we use are as follows.

(Q1) Insider attack and Cyber security in Internet of Things
(Q2) social network and text analysis and Internet of Things
(Q3) Security and privacy in Internet of Things
(Q4) Internet of Things in USA
(Q5) Big data and cloud computing and Internet of Things
(Q6) Distributed of Denial of Service Attacks with Internet of Things
(Q7) Cloud and Internet of Things and social network
(Q8) Attacks prevention in Internet of Things
(Q9) Types of algorithms in Internet of Things

TABLE II
PERFORMANCE TEST FOR PROPOSED SEARCH ENGINE.

Query	Hits	Time (ms)
(Q1)	29,490	40
(Q2)	44,390	39
(Q3)	24,100	35
(Q4)	11,500	20
(Q5)	65,480	65
(Q6)	17,700	25
(Q7)	58,600	42
(Q8)	38,500	30
(Q9)	10,700	20
(Q10)	80,900	75

(Q10) Business and Company Management system and Internet of Things

Note that some of these queries are more general than the others. Table II shows the number of relevant pages returned by the search engine and the times taken to find these pages. As can be seen by the table, the proposed search engine is fairly quick. Moreover, the amount or relevant documents are not exaggerated based on the amount of keywords. Queries with several keywords are generally the fastest since the result set is decreased extremely quickly. On the other hand, queries with particular keywords (there are no AND operators) are a little bit slower for the same reason. It can be noticed that the time is based on the amount of the initial result set.When a character search is running, the total list is iterated, and every URL is tested whether it verify the criteria.

After discussing its performance, we focus on computing the accuracy of the proposed search engine and comparing it with three general-purpose state-of-the-art search engines: Google, Yahoo and Bing. Needless to say, these search engines are developed and maintained by three of the largest corporations in the world employing some of the brightest minds and taking advantage of the latest technological advances and impressive infrastructures. Keeping this in mind, it would be impressive if our proposed search would perform comparably in terms of accuracy. Surprisingly, our proposed search engine noticeably outperform these search engines as shown later in this section. These results prove that even with simple implementations, specialized search engines are appealing as they outperform general purpose search engine.

For this experiment set, we use the same set of queries defined earlier in this section. We manually evaluate the results of each search engine to determine the percentage of relevant results. Since the number of returned results for each search engine massive, we limit our focus on the results shown in the first two pages (the top 20 URLs) to compute the system's accuracy.

Tables III, IV and V show the search results for the Google, Yahoo, Bing search engines, respectively, while Table VI shows the search results for the proposed search engine. Each of these tables show the percentage of relevant results for each query in addition to the index/rank of irrelevant results.

TABLE III
SEARCH RESULTS ON GOOGLE SEARCH ENGINE.

Query	Rel %	Index of non-Rel
(Q1)	40%	1,3,4,8,9,10,11,12,13,18,19,20
(Q2)	45%	9,10,11,13,14,15,16,17,18,19,20
(Q3)	100%	0
(Q4)	55%	5,6,7,9,10,13,17,19,20
(Q5)	100%	0
(Q6)	65%	5,7,8,10,11,14,17
(Q7)	85%	13,14,19
(Q8)	80%	2,11,17,20
(Q9)	95%	4
(Q10)	100%	0

TABLE IV
SEARCH RESULTS ON YAHOO SEARCH ENGINE.

Query	Rel %	Index of non-Rel
(Q1)	50%	1,3,6,7,11,13,15,16,17,18
(Q2)	40%	1,3,5,6,11,12,13,14,16,18,19,20
(Q3)	100%	0
(Q4)	75%	15,16,18,19,20
(Q5)	100%	0
(Q6)	35%	3,5,6,8,9,10,13,14,15,16,18,19,20
(Q7)	55%	4,6,8,13,14,15,17,19,20
(Q8)	60%	3,8,11,15,16,17,19,20
(Q9)	70%	2,9,10,17,19,20
(Q10)	100%	0

From these tables, one can easily see that all search engines work well for general queries, such as (Q3), (Q5) and (Q10) for which none of the returned results is irrelevant. On the other hand, specific queries, such as (Q1) and (Q2) are more difficult to handle. The tables show the superiority of Google search engine and our proposed search engine over Yahoo and Bing search engines. This is evident from the fact the for three queries the former two search engines provided the most accurate results whereas the latter two performed the best in only one query. Moreover, the average of relevant results for our proposed search engine is the highest (78.5%), closely

TABLE V
SEARCH RESULTS ON BING SEARCH ENGINE.

Query	Rel %	Index of non-Rel
(Q1)	55%	5,6,7,10,15,16,18,19,20
(Q2)	40%	2,3,4,5, 11,12,13,14,16,18,19,20
(Q3)	100%	0
(Q4)	55%	2,5,8,10,6,7,12,14,18
(Q5)	100%	0
(Q6)	35%	3,5,6,8,9,10,13,14,15,16,18,19,20
(Q7)	45%	1,2,3,5,7,10,11,13,16,19,20
(Q8)	60%	4,7,11,14,15,17,19,20
(Q9)	70%	3,8,10,13,19,20
(Q10)	100%	0

TABLE VI
SEARCH RESULTS ON PROPOSED SEARCH ENGINE.

Query	Rel %	Index of non-Rel
(Q1)	40%	9,10,11,12,13,14,15,16,17,18,19,20
(Q2)	50%	10,11,13,14,15,16,17,18,19,20
(Q3)	100%	0
(Q4)	70%	10,12,13,18,19,20
(Q5)	100%	0
(Q6)	60%	10,11,12,14,17,18,19,20
(Q7)	100%	0
(Q8)	80%	17,18,19,20
(Q9)	85%	18,19,20
(Q10)	100%	0

TABLE VII
RATIOS FROM SEARCH ENGINES.

Search engine	Avg Rel %
Google	76.5%
Yahoo	68.5%
Bing	66%
Proposed Engine	78.5%

followed by Google search engine (76.5%). On the other hand, the averages for Yahoo and Bing search engines are a bit far with 68.5% and 66%, respectively. Finally, in addition to the standard evaluation metric of relevancy, other evaluation metric exist such as duplication of some of the results, which is less of a problem for the proposed search engine than it is for other search engines.

V. CONCLUSION AND FUTURE WORK

The scale of Internet of Things (IoT) information on the Web is increasing exponentially which makes hard to discover the useful or suitable resources including informative knowledge and like articles. In the case of searching for common information resources, it is very difficult to determine the quality of results that come from general purpose search engines. In order to address such issues, we developed a topical search engine for IoT. In this paper, we described the overview of this search engine. We showed through experimentation that, even with simple implementation, the proposed search engine provided more accurate results than general purpose search engines developed and maintained by three of the largest corporations in the world. Improving the accuracy and performance of the proposed search engine is the topic of further research. Additionally, we intend to run it on powerful servers and allow large numbers of users to use it concurrently to evaluate its performance under real-world setting of high stress and massive demands.

REFERENCES

[1] D. Giusto, A. Lera, G. Morabito, and L. Atzori, *The Internet of Things*. Springer, 2010.

[2] M. Presser and A. Gluhak, "The internet of things: Connecting the real world with the digital world," *EURESCOM mess@ ge–The Magazine for Telecom Insiders*, vol. 2, 2009.

[3] A. Bassi and G. Horn, "Internet of things in 2020: A roadmap for the future," *European Commission: Information Society and Media*, 2008.

[4] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[5] D. Evans, "The internet of things: how the next evolution of the internet is changing everything," Cisco Internet Business Solutions Group (IBSG), White paper, April 2011.

[6] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica, "A data-oriented (and beyond) network architecture," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 4, pp. 181–192, 2007.

[7] W. contributors. (2014, September) Internet of things. [Online]. Available: http://en.wikipedia.org/wiki/Internet_of_Things

[8] I. Toma, E. Simperl, and G. Hench, "A joint roadmap for semantic technologies and the internet of things," in *Proceedings of the Third STI Roadmapping Workshop, Crete, Greece*, vol. 1, 2009.

[9] A. Katasonov, O. Kaykova, O. Khriyenko, S. Nikitin, and V. Y. Terziyan, "Smart semantic middleware for the internet of things." *ICINCO-ICSO*, vol. 8, pp. 169–178, 2008.

[10] W. Wahlster, "Web 3.0: Semantic technologies for the internet of services and of things," Lecture at the 2008 Dresden Future Forum, June 2008.

[11] I. Vázquez, "Social devices: Semantic technology for the internet of things," Week@ ESI, Zamudio, Spain, June 2009.

[12] D. Guinard and V. Trifa, "Towards the web of things: Web mashups for embedded devices," in *Workshop on Mashups, Enterprise Mashups and Lightweight Composition on the Web (MEM 2009), in proceedings of WWW (International World Wide Web Conferences), Madrid, Spain*, 2009, p. 15.

[13] T. OReilly and J. Pahlka, "The web squaredera," *Forbes*, September 2009.

[14] G. Broll, E. Rukzio, M. Paolucci, M. Wagner, A. Schmidt, and H. Hussmann, "Perci: Pervasive service interaction with the internet of things," *Internet Computing, IEEE*, vol. 13, no. 6, pp. 74–81, 2009.

[15] A. Ilic, T. Staake, and E. Fleisch, "Using sensor information to reduce the carbon footprint of perishable goods," *IEEE Pervasive Computing*, vol. 8, no. 1, pp. 22–29, 2009.

[16] A. Dada and F. Thiesse, "Sensor applications in the supply chain: the example of quality-based issuing of perishables," in *The Internet of Things*. Springer, 2008, pp. 140–154.

[17] A. Vilamovska, E. Hattziandreu, R. Schindler, C. Van Oranje, H. De Vries, and J. Krapelse, "Rfid application in healthcare–scoping and identifying areas for rfid deployment in healthcare delivery," *RAND Europe*, February 2009.

[18] M. Revati, K. N. Rao, M. K. Babu, K. Ramakrishna, C. R. Jacob *et al.*, "A novel search engine to trace medical information needs using medical domain ontology," *International Journal on Computer Science & Engineering*, vol. 3, no. 8, 2011.

[19] D. Tumer, M. A. Shah, and Y. Bitirim, "An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia," in *Internet Monitoring and Protection, 2009. ICIMP'09. Fourth International Conference on*. IEEE, 2009, pp. 51–55.

[20] G. Madhu, D. A. Govardhan, and D. T. Rajinikanth, "Intelligent semantic web search engines: A brief survey," *arXiv preprint arXiv:1102.0831*, 2011.

[21] H. Dietze and M. Schroeder, "Goweb: a semantic search engine for the life science web," *BMC bioinformatics*, vol. 10, no. Suppl 10, p. S7, 2009.

[22] P. De Bra, G.-J. Houben, Y. Kornatzky, and R. Post, "Information retrieval in distributed hypertexts." in *RIAO*, 1994, pp. 481–493.

[23] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through url ordering," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 161–172, 1998.

[24] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific web resource discovery," *Computer Networks*, vol. 31, no. 11, pp. 1623–1640, 1999.

[25] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Building domain-specific search engines with machine learning techniques," in *AAAI Spring Symposium on Intelligent Agents in Cyberspace 1999*, 1999.

[26] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, M. Gori *et al.*, "Focused crawling using context graphs." in *VLDB*, 2000, pp. 527–534.

[27] P. Chahal, M. Singh, and S. Kumar, "Ranking of web documents using semantic similarity," in *Information Systems and Computer Networks (ISCON), 2013 International Conference on*. IEEE, 2013, pp. 145–150.

[28] Y.-h. Feng, Y. Hong, W. Tang, J.-m. Yao, and Q.-m. Zhu, "Using html tags to improve parallel resources extraction," in *Asian Language Processing (IALP), 2011 International Conference on*. IEEE, 2011, pp. 255–259.

[29] A. Pesaranghader, A. Pesaranghader, N. Mustapha, and N. M. Sharef, "Improving multi-term topics focused crawling by introducing term frequency-information content (tf-ic) measure," in *Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on*. IEEE, 2013, pp. 102–106.

[30] U. Gurav and R. Shaikh, "Virtualization: a key feature of cloud computing," in *Proceedings of the International Conference and Workshop on Emerging Trends in Technology*. ACM, 2010, pp. 227–229.

[31] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.