TOWARDS INFORMATION RETRIEVAL CHARACTERISTICS IN THE INTERNET OF THINGS

A. Andrushevich, A. Kurbatski, V. Basko, A. Klapproth

*Software Engineering Department, Faculty of Applied Mathematics
and Computer Science, Belarusian State University
Minsk, Belarus
E-mail: andrushevich@bsu.by*

This work describes an initial research for expanding the common characteristics of ***Information Retrieval*** (IR) to the appearing ***Internet of Things*** (IoT). Starting with latest advances of IoT the authors briefly sketch several examples when IR can become beneficial for end-users. Next, the article gives an analytical interpretation of typical IoT characteristics seen in the recent developments. Resting on common IR procedure and widely used IR metrics as well as using previously given IoT characteristics the authors derive several basic characteristics for IR in IoT. Finally, the paper is finalized by conclusion and outlook.

Key words: Internet of Things, Big Data, Information Retrieval, Context-awareness.

Introduction

Dynamic developments of recent years in information and communication technologies rapidly bring visionary concept of ***Internet of Things*** (IoT) [1] to the real life. Initially starting from specialized business branches like logistics and industries like supply chain management in the late 90-ies the IoT is quickly approaching broad end-user markets like consumable electronics nowadays. Main visionary feature in this regard is the interconnection and interoperability between previously separated application domains like traditional building automation, safety and security systems, intelligent transportation systems and weather forecasting to name a few. Technologically this trend is enabled on one hand through wide adoption of standardized and / or proprietary ***Wireless Sensor Networks*** (WSNs) and on the other hand recent release and penetration of low power embedded implementations of IPv6 stack, namely 6LoWPAN, to build an IPv6-compatible WSNs. These two communication technologies are already starting and will be increasingly uniting different application domains into one IoT ecosystem. Furthermore, mobile and intelligent ***Human-Computer Interaction*** (HCI) [2] platforms (smart phones, tablets, other personal mobile devices) are becoming more popular for variety of tasks like monitoring, controlling and communicating with geographically distributed physical objects, empowered with IoT object connectivity platforms. These objects start from household and consumer devices (TV-set, dishwasher, vacuum cleaner, etc.) within users' proximity, where additional information and control via ***Augmented Reality*** (AR) makes sense, and span to parked somewhere personal vehicles, requiring for example payment for additional parking time, or even to summer / winter vacation houses equipped with remote building automation components (access control systems, HVAC, maintenance, etc.), building ensembles and districts of future smart cities. At the same time, based on recognized user needs and intentions semantically relevant assisting suggestions [3] could be delivered to the user here and now especially in emergency cases, or even when simply hopping through a new unknown

public place. Embedded devices for implementation of human-machine interaction vary from smart phones to newly developed AR systems based on **Head-Mounted Displays** (HMDs).

The rest of this article focuses on creating the factual and theoretic basis for identifying particular characteristics of **Information Retrieval** (IR) in the IoT. In chapter 2 authors analytically interpret a set of typical IoT characteristics that are becoming known from appearing best-practices and concepts for IoT deployments. In chapter 3 the authors are focused on describing classic procedure used in the information retrieval process and also some common metrics used for performance measurement of IR systems. Chapter 4 is devoted to the theoretical derivation of some basic characteristics of IR in the IoT. Chapter 5 concludes the paper and points out directions for the future research.

Typical iot characteristics

Numerous applications, recent intensive developments of IoT technologies and al-ready started IoT pilot deployments allowed several research and development com-munities to identify and describe some common IoT characteristics and patterns [4]. According to our opinion, however, one of the most representative, well-generalized and application independent overview over IoT characteristics has been proposed by [5, 6] in so called paradigm of **Fog Computing** (FC) that locally extends globally centralized **Cloud Computing** (CC) to numerous bound to the place end devices via the Edge of the Network notion with its following crucial necessary features:

- Heterogeneity based on existence of different technologies, form factors and variety of deployment environments
- Interoperability between heterogeneous technologies services and applications
- Low – latency multimedia traffic from video streaming, gaming and AR applications and services running at becoming more and more powerful consumer devices
- Real-time orchestration, cooperation and analytics between involved applications and services as well as real-time interactions with end-users
- Geographically distributed computational infrastructure deployments to run high-quality multimedia **Location-Based Services** (LBS)
- Very high system scalability with huge number of nodes (up-to hundreds of millions end-devices) based on multi-tiered hierarchically organized distributed com-putting / storage resources as a consequence of the wide geographic distribution
- Mobility support for delivering reliable services between moving objects, often implying decoupled host and location identity; requiring a distributed directory system
- Predominance of wireless communications
- Basic analytics for data processing close to the source like fusion from multiple sources and basic filtering
- Interface to CC platforms in regard of **Big Data** (BD) processing like filtered data aggregation from the fusion nodes

The phenomenon of BD has been mainly associated with CC services until now. But further consequences of wide IoT adoption are also enormous amounts of data automatically produced by billions of devices and sensors. Part of this data like **Machine-to-Machine** (M2M) interaction, data collection and actuator control command requires real-time processing [5]. The other part like data analytics in form of reporting or visualization dashboards is less time-critical but it requires highly reliable, large and near to permanent data storage infrastructure. BD phenomenon has been commonly associated with 3 features: data volume, velocity and variety [7]. Based on the described above characteristics one can see

that wide deployment of IoT will not only impact all features but will also add a new one, namely *geographic distribution* of stored data between IoT layers [5].

Vision enabled by ubiquitous context-aware IoT [8] is generally covering the technological and societal mega-trends of informational assistance content personalization enabled via human activity / behavior monitoring and on-demand service provision enabled via growing technological connectivity of physical objects and devices. Using modern data gathering technologies (RFID, NFC, WSN, etc.) it is possible and rational to associate various real-world entities with personalized dynamic content. For example, public transportation passengers massively make last-minute seat reservations during rush hours; commuters want to save time and fuel by purchasing consumables on their way home; and shop customers search contextualized information about demanded goods and services from relevant sources such as competitors (for price), government agencies (for safety) and friends (for quality). It is clear that the scale of specialized and heterogeneous demands fluctuates with time, location and circumstances. Furthermore, lots of demands are unexpressed or get lost. To address these issues, a context-aware IoT system can bring an unprecedented level of harmony between users and their needs, as well as consumers and producers via gaining the near-to-real-time feedback and control on resources consumption and production [9].

Finally, even if IoT is mainly about machine-to-machine communications, all the enabled via cloud and fog computing [5, 6] mechanisms, processes and algorithms of dealing with context and location awareness, security and human behavior predicting are not the main goal of itself but are only means to deliver an additional value to end-users.

Information retrieval

Classic application examples of information retrieval are traditionally focused on searching methods and mechanisms of looking up for information within files and for metadata about files, as well as within and about databases, the web documents, and other sources. Information retrieval is also often described as a process of identifying a set of documents that are relevant to selected topic or satisfy to some query [10]. It is important to mention that IR is about finding the existing information and not about discovering new knowledge like it is done in data mining.

Information retrieval process includes the following often iteratively repeated steps:
1. Information need definition and formal statement of information query
2. Identification of all possible information databases (retrieval sources)
3. Information extraction from the identified data sources
4. Studying the extracted information and retrieval results assessment

The situation is trivial when IR addresses concrete information need for some fact. But as soon as problem-oriented information need considered, it is often not enough to perform a search using one query in one source database [11]. Instead, one has to request multiple search databases while constantly updating the query. The problem of access to relevant complex data is on one side caused by vagueness and ambiguity of natural language used to express targeted keywords and queries. Further, the search databases often contain uncertain or insufficient metadata describing contents of entries. That is why the IR query also rarely uniquely identifies a single entry in search database. Instead, several entries usually match the query with different degrees of relevancy. Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value [12]. In case of elementary example like a web-page with text and images, a web crawler preliminary departs at a trusted starting point and hops along all hyperlinks met on this web-page. For each hyperlink the crawler determines the credibility and

link value by use of a page ranking algorithm like HITS. All this credibility data is then been used in assigning the ranking value to the objects in search database [13].

Considering commonly used metrics for measuring the performance of IR algorithms requires a ground truth notion of **Relevancy**. If *B* is a number of relevant entries returned by IR algorithm *A* for query *Q*, *C* is a number of non-relevant entries returned by IR algorithm *A* for query *Q*, and *D* is a number of relevant entries not returned by IR algorithm *A* for query *Q*, then precision *P* of IR algorithm *A* for query *Q* is defined as $P = \dfrac{B}{B+C}$, and recall *R* of IR algorithm for query *Q* is defined as $R = \dfrac{B}{B+D}$. Additionally, the weighted harmonic mean of precision and recall is used for IR performance measurement: $F_1 = \dfrac{2 \cdot P \cdot R}{P+R}$, that is a case if β=1 of traditional *F*-measure $F_\beta = \dfrac{(1+\beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R}$.

For problem-oriented IR where a ranked sequence of entries is returned, it is often useful to see the sorting order of returned entries. If one computes precision and recall for every ranked entry in the returned sequence it is also possible to associate values for precision *P(R)* as a function of recall *R*. In this case one can also calculate the average precision [14] as average value of *P(R)* over the interval, where *R* changes from 0 to 1: $\overline{P} = \int_0^1 P(R)dR$. In practice this integral is replaced with a finite sum over every position in the ranked sequence of entries.

Some characteristics of ir in iot

Described above general IR procedure when users actively search for information are called pull-services [15, 16]. Push-services instead provide users with information based on subscriptions and are implemented in form of notifications when new relevant information appears. One can see that push-services are well-suited for implementation of context-awareness in IoT, serving the characteristics of real-time and low-latency applications. Further, the ideal IR result in IoT consists not only of correct relevant knowledge in correct representation form with acceptable quality but also of *appropriate time and geographical delivery place* of information results. On the other side the IoT contributes to IR via extension of previously hardly or not possible to digitalize data from physical environments or even from some biological and chemical processes.

Because of IoT characteristics of real-time, low-latency, geographic distribution and mobility support described in Chapter 2, it is logical to foresee that amount and contents of relevant entries in search databases can change significantly depending on geographic place where from and depending on time when the same location-based and possibly implemented as a push-service query *Q* is being performed. Hence, numbers of relevant entries returned by IR algorithm *A* as well the number of relevant entries not returned by IR algorithm *A* for query *Q* can be considered as functions of geographical position and time *B(x, y, z, t)*, *C(x, y, z, t)*, *D(x, y, z, t)* or if *s* is a coordinate *(x, y, z)*, then *B(s, t)*, *C(s, t)*, *D(s, t)*. In this case precision, recall, weighted harmonic mean, and average precision will also become functions of geographical position and time $P(s, t)$, $R(s, t)$, $F_1(s, t)$, $\overline{P}(s,t)$.

Furthermore, assuming that average precision is a continuous function of time and space in some IoT location and time-based services and applications it is of benefit to cal-

culate average precisions over the time interval between t_0 to t_1 as $\overline{\overline{P}}_t(s) = \int_{t0}^{t1} \overline{P}(s,t)dt$ and over the cuboid between $s_0(x_0, y_0, z_0)$ to $s_1(x_1, y_1, z_1)$ as

$$\overline{\overline{P}}_s(t) = \int_{s0}^{s1} \overline{P}(s,t)ds = \iiint_S \overline{P}(x, y, z, t)dxdydz .$$

These average values can be for example used to estimate the Quality of Services, personalization and context-awareness levels in particular time intervals and within particular service regions.

Conclusion and outlook

Looking at rapidly developing IoT and Human Machine Interaction research it is becoming clear that real and virtual worlds will be eventually merged resulting in applications like digital search in the real world (mainly characterized by time and space) and natural search experiences while looking for virtual items (mainly characterized by information properties). In the described above research we contribute to understanding the nature of this merge by creating initial factual and theoretic basis for identification and derivation of particular characteristics of IR in the IoT. However, one has to also admit that derivation and formal description of physics-digital laws describing how both worlds meet each other is only at initial stage nowadays. Hence, authors expect intensification of this research area through new theoretic hypothesis as well as different empiric studies including but not limited to computational experiments, simulations, benchmarking of appearing IoT pilots, etc.

Bibliography

1. *Kurakova T.* Overview of Internet of Things // Proceedings of the INTHITEN (INternet of THings and ITs ENablers) conference, Saint Petersburg, Russia, 3–4 June 2013, P. 1–13.
2. *Jacko J.* The Human-Computer Interaction Handbook // (3rd ed.) CRC Press, 2012.
3. *Tsetsos V., Anagnostopoulos C., Kikiras P., Hadjiefthymiades S.* Semantically enriched navigation for indoor environments // Int. Journal of Web/Grid Services, December 2006, V. 2. P. 453–478.
4. *Arzori L., Iera A., Morabito G.* The Internet of Things: A Survey // Computer Networks. 2010. V. 54. №. 15. P. 2787–2805.
5. *Nemirovsky M., Milito R., Yanuzzi M.* Fog computing // Class Conference ICREA/BSC, 24.10.2012.
6. *Bonomi F., Milito R., Zhu J., Addepalli S.* Fog Computing and Its Role in the Internet of Things // Proceedings of the 1st edition of the MCC workshop on Mobile cloud computing, 2012. P. 13–16.
7. *Ebbers M., Abdel-Gayed A., Budhi V., Dolot F., Kamat V., Picone R., Trevelin J.* Addressing Data Volume, Velocity, and Variety with IBM InfoSphere Streams V3.0 // IBM Redbooks, 12.03.2013.
8. *Preuveneers D., Berbers Y.* Internet of Things: A Context-Awareness Perspective // Book chapter in "The Internet Of Things: From RFID to the Next-Generation Pervasive Networked Systems", edited by L. Yan et al. Auerbach Publications, Taylor & Francis Group, USA, March, 2008.
9. BUTLER – uBiquitous, secUre inTernet-of-things with Location and contEx-awaReness. Research project of the European 7th Framework Programme, http://www.iot-butler.eu/
10. *Baeza-Yates R., Ribeiro-Neto B.* Modern Information Retrieval // ACM Es, 1999.
11. *Bates M.* The Design of Browsing and Berrypicking Techniques for the Online Search Interface. 1989.
12. *Frakes W.* Information Retrieval Data Structures & Algorithms // Prentice Inc, 1992.
13. *Manning C., Raghavan P., Schütze H.* Introduction to Information Retrieval // Cambridge University Press, 2008.
14. *Zhu M.* Recall, Precision and Average Precision // Working Paper, Department of Statistics & Actuarial Science, University of Waterloo, 26.08.2004.
15. *Poetzsch E.* Information Retrieval – Einführung in Grundlagen und Methoden // E. Poetzsch Verlag. Berlin, 2006.

16. *Stock W. G.* Information Retrieval. Informationen suchen und finden. München, Wien: Oldenbourg, 2007.