International Journal of Control and Automation
Vol. 11, No. 5 (2018), pp.173-182
http//dx.doi.org/10.14257/ijca.2018.11.5.15

A Method of Trend Analysis using Latent Dirichlet Allocation

Myeong-Ha Hwang[1], Suwook Ha[2], Minkyo In[3] and Kangchan Lee[4*]

[1]University of Science and Technology (UST)
[2,3,4]Electronics and Telecommunications Research Institute (ETRI)
[1]raphael9290@gmail.com, [2]sw.ha@etri.re.kr, [3]mkin@etri.re.kr, [4]chan@etri.re.kr*

Abstract

Due to the introduction of text mining, studies have been conducted to analyze meaningful topics and trends in document collections. Trend analysis using Latent Dirichlet Allocation (LDA) in text mining is adopted as one of the trend analysis methods with high accuracy. In this paper, we propose a trend analysis method using LDA. The method is composed of 5 steps and the trend analysis is performed by topic using the extracted result combining LDA and Top 10 keywords. By applying our method and LDA to international standards documents, we performed topic modeling and checked the trend of international standards using extracted topics.

Keywords: *Trend Analysis, Text Mining, Latent Dirichlet Allocation*

1. Introduction

Recently, studies have been conducted to analyze meaningful topics and trends in document collections. Trend analysis is defined as a method of projecting the appearance of future social changes based on current and historical data. The trend analysis method that was mainly used in the past is a qualitative trend analysis. Qualitative trend analysis has a probability that involve individual subjectivity by using opinion of experts. On the other hand, quantitative trend analysis has no probability of subjectivity. Therefore, quantitative trend analysis has been introduced to overcome limitations of qualitative trend analysis. Text mining, which is one of the quantitative trend analysis, is possible to extract topics from accumulated papers. Representative topic modeling algorithm in text mining is LDA. The LDA is an algorithm that overcome the disadvantages of the Probabilistic Latent Semantic Indexing (PLSI) and finds hidden topics in the documents as a generation model. However, there is a lack of research on methods to analyze trends using LDA.

In this study, we propose a method of trend analysis using occurrence rate of each keyword in extracted topic by LDA. The proposed method is designed in 5 steps. We use 3,962 International Telecommunication Union Telecommunication Standardization Sector (ITU-T) recommendations and extract representative topics (Top 10 keywords) from each document using LDA and construct them as a topic table. The proposed method is applied to extract the trends of topics after the user chooses a range of documents. The trend graph shows the sum of the occurrence rate of each topic by year. Furthermore, we find the keywords that affect the drawing of an upward graph within each topic.

2. Related Work

A trend is defined as a method of describing changes over a long period of time and can predict future using past patterns[1]. Qualitative trend analysis uses the opinions of experts to extract trends in each period and then analyzes the trends by comparing the

Received (October 30, 2017), Review Result (February 6, 2018), Accepted (April 9, 2018)

[*] Corresponding Author

ISSN: 2005-4297 IJCA

similarities of the trends[2]. Interval-Halving algorithm is used to extract trends and time-weighted average is calculated to compare the similarity of two trends[3]. However, the qualitative trend analysis has the probability that individual subjectivity will intervene. Therefore, studies are underway to overcome these limitations by introducing quantitative trend analysis[4].

Quantitative trend analysis is a method of trend analysis based on actual data and literature. One of the most popular and accurate trend analysis methods is the trend analysis method using text mining. In particular, the researches of trend analysis method using extracted topic using LDA has been conducted.

LDA is a generation model that finds hidden topics in documents. The generation model is considered to be the process of creating the document. Therefore, hidden variable such as document structure is inferred through observed variable such as document, word and so on. It is possible to find out the topic of the entire document set, the probability of topic of each document, and the probability that each word is included in each topic. It can be expressed as a probabilistic graph model as shown in Figure 1[5].

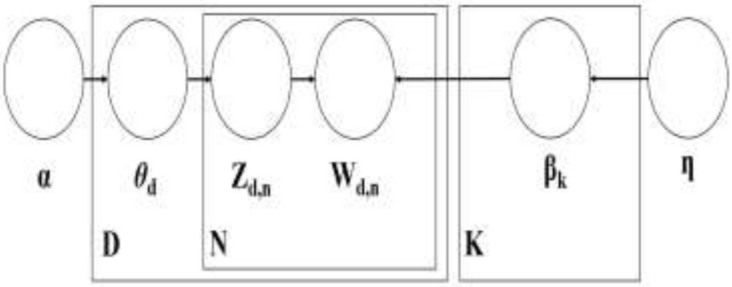

Figure 1. Probabilistic Graph Model of Latent Dirichlet Allocation

The N plate denotes the set of words and the D plate denotes the set of topics. The K plate denotes the number of clusters of extracted topic. Each node is labeled according to its role in the generative model process and is a random variable. The LDA can predict hidden variables such as topic(β), per-word topic assignment(Z) and topic proportions(θ) using observed variables(W). These parameters are extracted based on the dirichlet parameters(α ,η). The dirichlet parameters are the word counts including each topic.

The first step for the topic extraction and trend analysis method using LDA is to extract topics using LDA from given set of documents. The second step is to perform the trend analysis using the frequency of the topics shown in each interval after dividing the extracted topics into a period of time. This trend analysis method has been transformed into various forms such as TOT(Topics over Time) and On-line Trend Analysis with Topic Models[6][7]. TOT is an algorithm that can recognize the time series by topic using LDA. On-line Trend Analysis with Topic Models help people understand the trend of Twitter by using On-line LDA(OLDA).

3. Proposed Method

3.1. Data Set

In order to apply the trend analysis method proposed in this study, we have collected 3,962 ITU-T recommendations on July 27, 2017 from the ITU official website[8]. The ITU-T recommendations consist 22 series from D series to Z series. The series is categorized into a set of international standard documents related to each subject as per Table 1.

Table 1. Series of ITU-T Recommendations

Series	Subjects	The number of documents
D	General tariff principles	112
E	Overall network operation, telephone service, service operation and human factors	235
F	Non-telephone telecommunication services	137
G	Transmission systems and media, digital systems and networks	468
H	Audiovisual and multimedia systems	303
I	Integrated services digital network	167
J	Cable networks and transmission of television, sound programme and other multimedia signals	87
K	Protection against interference	105
L	Environment and ICTs, climate change, e-waste, energy efficiency, construction, installation and protection of cables and other elements of outside plant	130
M	Telecommunication management, including TMN and network maintenance	222
N	Maintenance: international sound programme and television transmission circuits	26
O	Specifications of measuring equipment	39
P	Terminals and subjective and objective assessment methods	94
Q	Switching and signaling	696
R	Telegraph transmission	71
S	Telegraph services terminal equipment	31
T	Terminals for telematic services	127
U	Telegraph switching	49
V	Data communication over the telephone network	78
X	Data networks, open system communications and security	473
Y	Global information infrastructure, internet protocol aspects and next-generation networks	252
Z	Languages and general software aspects for telecommunication systems	60

3.2. A Design of Structure of Proposed Method

The proposed method of trend analysis in this study consists of 5 steps as shown in Figure 2 and the explanation of each step is as follows.

(1) Construct the topic tables

- Extract keywords of top 10 after applying wordcount for each document

- Normalize wordcount of extracted keywords to extract occurrence rate of word

(2) Results of Topic Modeling

- Configurate the scope of documents for topic modeling

- Implement the LDA after setting the number of clusters and the number of iterations

(3) Extract the table of keyword in topic

- Normalize the dirichlet parameter of each keyword to extract occurrence rate of LDA

(4) Normalize the occurrence rate by keyword

- Normalize the occurrence rate by year for keywords of top 10 in a topic

(5) Extract the graph of trend by topic

- Sum normalized occurrence rate by year

- Extract graph with summed values

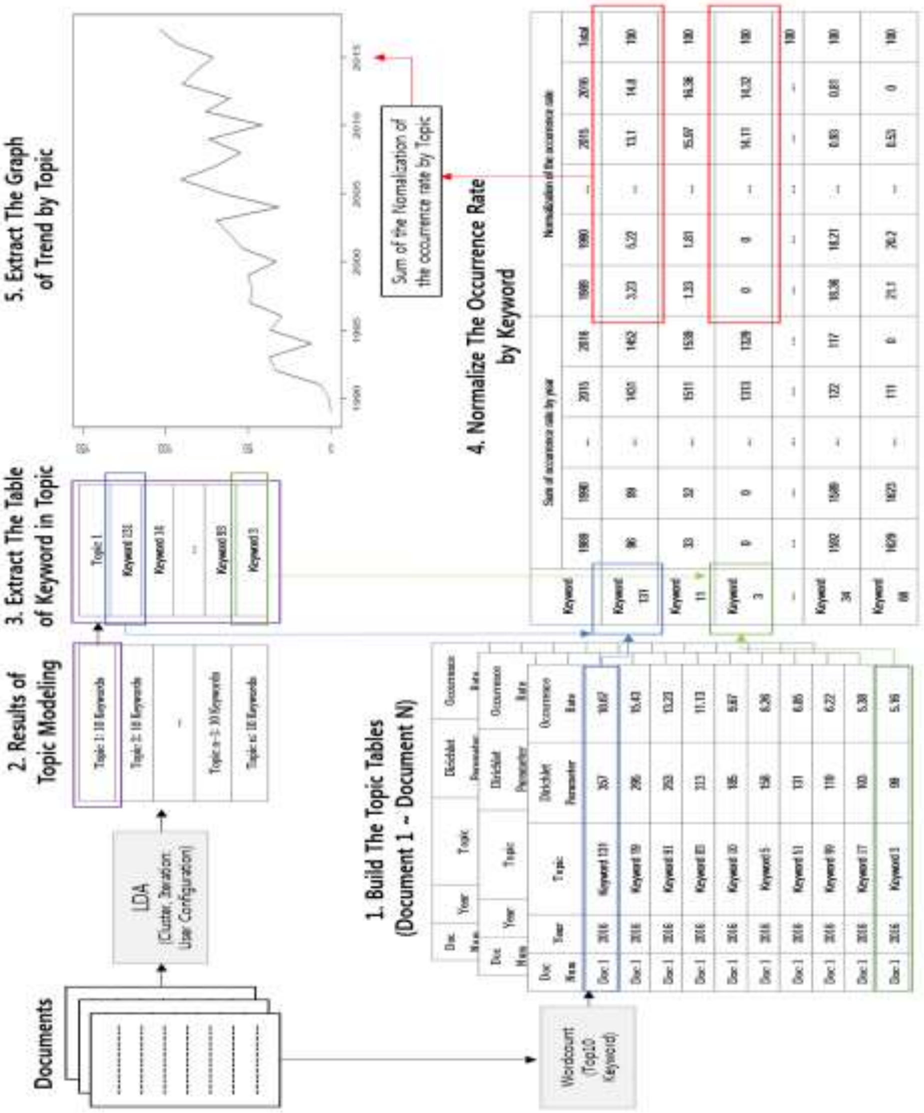
Documents
LDA (Cluster, Iteration, User Configuration)
Wordcount (Top10 Keyword)
2. Results of Topic Modeling
Topic 1: 10 Keywords
Topic 2: 10 Keywords
...
Topic n-1: 10 Keywords
Topic n: 10 Keywords
3. Extract The Table of Keyword in Topic
Topic 1
Keyword 131
Keyword 14
...
Keyword 93
Keyword 3
1. Build The Topic Tables (Document 1 ~ Document N)
Doc Num
Year
Topic
Dirichlet Parameter
Occurrence Rate
4. Normalize The Occurrence Rate by Keyword
Sum of occurrence rate by year
Normalization of the occurrence rate
Keyword
Total
5. Extract The Graph of Trend by Topic
Sum of the Normalization of the occurrence rate by Topic

Figure 2. Structure of Proposed Method of Trend Analysis

4. Experiments and Results

4.1. Results of Topic Modeling by Period

We performed periodic topic modeling with 4-year period using LDA, and then selected 5 representative topics for each period as shown in Table 2.

Table 2. Topic Modeling Results by Each Period (Series: All, Period: 4 Year, The Number of Topic: 20, The Number of Iteration: 1,000)

Period	Topic Name	Topic
2016 – 2013	Topic 1	Service, Network, Cloud, Information, Application, Device, User, Capability, Support, Function
	Topic 2	ONU, OLT, PON, Message, Upstream, ID, Frame, Downstream, Bit, Byte
	Topic 3	Security, Service, Information, Authentication, User, System, Datum, Key, Management, Access
	Topic 4	Stream, Video, Object, IPTV, System, Content, Define, Information, Field, Downstream
	Topic 5	Energy, ICT, Datum, System, Network, Impact, Change, City, Service, Climate
2012 – 2009	Topic 1	Network, Service, Function, NGN, Control, Information, Application, Access, User, Telecommunication
	Topic 2	Service, Content, IPTV, Key, User, Authentication, Function, Datum, Information, Message
	Topic 3	Video, Frame, Terminal, Parameter, Packet, Channel, Audio, Rate, Quality, Signal
	Topic 4	Connection, Call, Network, Layer, Packet, Route, Link, Control, Plane, Interface
	Topic 5	Bit, NPAR, Octet, TC, TPS, Downstream, RFC, SIP, TS, Upstream
2008 – 2005	Topic 1	Network, Service, Layer, System, Control, Information, Time, Packet, QoS, model
	Topic 2	TS, ATIS, WWW, HTTP, Org, Service, CCSA, 3G, Approve, TTC
	Topic 3	Security, Key, Authentication, Service, Network, User, System, Datum, Information, Access
	Topic 4	BIP, Value, Component, BioAPI, Function, Message, Parameter, EndPoint, RFC, Send
	Topic 5	Optical, System, Test, Cable, Fibre, Power, db, Signal, Line, Function
2004 – 2001	Topic 1	Service, Network, System, Transmission, Telephone, Circuit, Telecommunication, International, Cable, Signal
	Topic 2	HTTP, WWW, TS, CWTS, Service, Org, TTC, Publish, Protocol, Internet
	Topic 3	Cell, Frame, Connection, Packet, ATM, Network, Bit, Field, Delay, Parameter
	Topic 4	SCF, Operation, Call, Parameter, Error, Service, Procedure, User, Event, Message
	Topic 5	Route, Network, Traffic, Link, Connection, Bandwidth, Call, Path, Method, Capacity
2000 – 1997	Topic 1	Network, Service, Route, Circuit, System, Connection, Transmission, Traffic, International, Time
	Topic 2	Call, Parameter, Message, SCF, DP, Information, Connection, Signal, SSF, Network
	Topic 3	Bit, Signal, Datum, Modem, Transmit, Channel, db, DTE, Echo, ATU
	Topic 4	TP, Dialogue, Commit, Request, Indication, Rollback, Transaction, Issue, IND, AF
	Topic 5	Cell, Function, ATM, Connection, Layer, Network, VC, Signal, Bit, VP
1996 – 1993	Topic 1	Service, Network, ISDN, Function, Connection, Access, User, Information, Terminal, Interface
	Topic 2	Call, Packet, DTE, Frame, Bit, Address, Datum, Facility, Message, User
	Topic 3	Security, Service, Information, Access, Datum, Application, Control, Entity, Authentication, System
	Topic 4	System, Circuit, Level, Value, db, Measurement, Test, Signal, Loss, Bit
	Topic 5	Signal, Call, Service, Telex, Message, International, Terminal, Station, Network, User
1992 – 1989	Topic 1	Service, Management, Network, Layer, Security, System, Datum, Mechanism, Information, Entity
	Topic 2	Call, User, Service, Supplementary, Network, Operation, Party, Request, Charge, Impact
	Topic 3	DTAM, Primitive, Parameter, Service, Transfer, Request, User, PM, Indication, Capability
	Topic 4	POC, User, Service, Stop, Terminal, Trigger, Datum, Session, QoS, Request
	Topic 5	DTE, Call, Signal, Telex, DCE, Information, Character, Network, Datum, Parameter

4.2 Results of Trend Analysis

We applied proposed method to analyze the trends of international standards. Trend graph is shown in Figure 3. The graph is based on the occurrence rate of each topic using our method. Looking at the trend graph according to the occurrence rate of each topic in the period, it can be seen that the graph has been upward from 2013 to 2016. If we extract a trend graph for a topic extracted in other periods, we can see that the graph shows upward in the selected period.

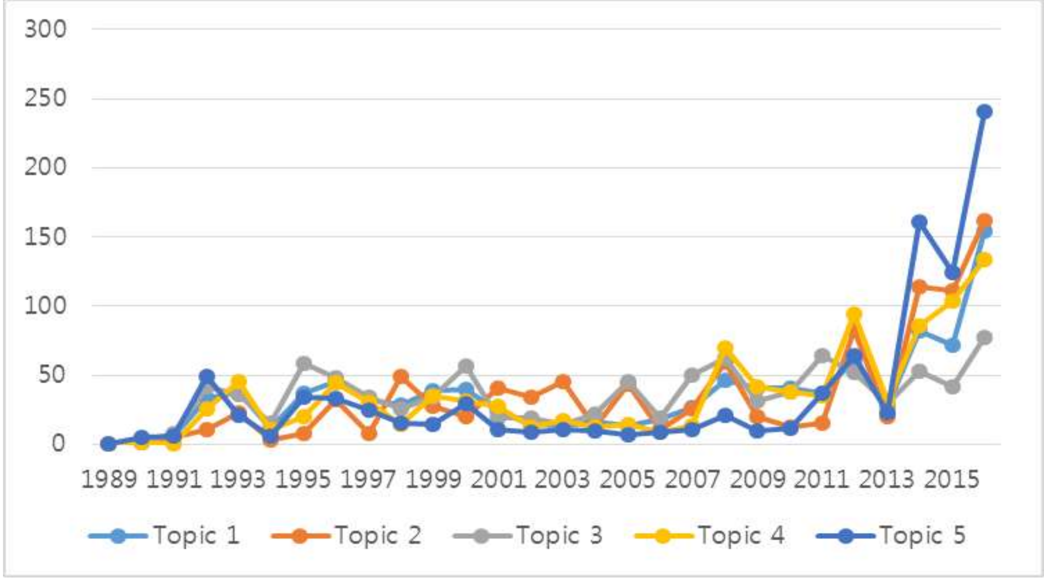

Figure 3. Trend Graph based on Occurrence Rate of Each Topic (2013 – 2016)

In order to examine the keywords that affect the trend, we extract the trend graph based on each keyword in topic as shown in Figure 4. We set the period from 2013 to 2016 to understand recent trends. The results of this experiment show there are keywords that affect the trend in topic.

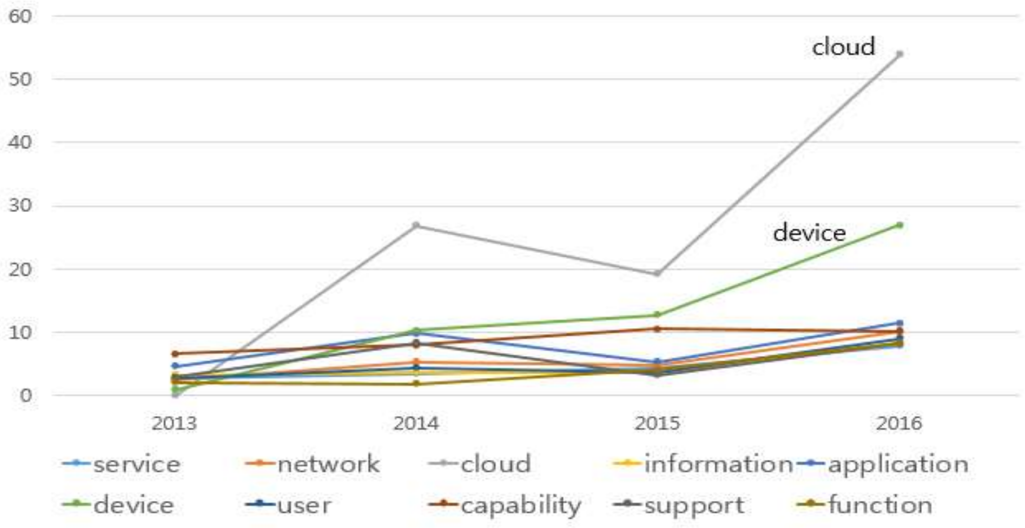

Figure 4. Trend Graph based on Each Keyword in Topic 1 (2013 – 2016)

In order to check the upward keyword in the topic based on the trend graph, the slope of each section of the graph for each keyword in the topic 1 is calculated as shown in Table 3. The order of the upward keywords is selected through the sum of the slopes of each section. As a result, it can be confirmed that ‘Cloud’ and ‘Device’ keywords are representative upward keywords in topic.

Table 3. Keyword Slope of Topic 1 (2013 - 2016)

Keyword	2013-2014	2014-2015	2015-2016	Total
service	0.78	0.97	3.48	5.23
network	2.81	-0.44	5.25	7.62
cloud	26.85	-7.68	34.82	53.99
information	0.4	0.11	4.6	5.11
application	5.13	-4.55	6.21	6.79
device	9.41	2.46	14.25	26.12
user	1.68	-0.79	5.24	6.13
capability	1.37	2.64	-0.43	3.58
support	5.51	-5.11	4.95	5.35
function	-0.24	2.12	4.3	6.18

As shown in Table 4, two representative keywords in each topic are extracted. In topic1, ‘cloud’ and ‘device’ are keywords that represent the period, as selected in Gartner Trends 2013, 2014 and 2015 [9-11]. According to Gartner Trends, cloud-based trends such as hybrid cloud and personal cloud are changing every year. Next, we confirmed that ‘security’ in topic 4 and ‘byte’ in topic 3 issues have emerged during this period as the ‘cloud’ related technology developed. It is natural that the importance of transmission units and security system for cloud service are emerging [12, 13]. ‘video’ in topic 2 and ‘city’ in topic 5 keywords are related to Machine Learning and Internet of Things(IoT). As the development of Machine Learning and IoT technology progressed rapidly during the period, the researchers who wanted to build a smart city were active [14].

Table 4. Results of Hot Keywords Extraction for Each Topic (2013 - 2016)

Topic Name	Hot Keywords
Topic 1	Cloud, Device
Topic 2	Video, Downstream
Topic 3	Byte, ID
Topic 4	System, Security
Topic 5	City, Climate

5. Conclusion

In this paper, we proposed a method of trend analysis using text mining and discussed the proposed method in each step such as construction of the topic tables, results of topic modeling, extraction of the keyword table in topic, normalization of the occurrence rate by keyword, extraction of the graph of trend by topic. We extracted representative topics for each 4-year period and the trend graphs of each topic using ITU-T recommendations. In addition, we confirmed hot keywords representing the recent period by extracting the trend graph of the keywords included in each representative topic.

As a results of the experiments, we found out that the technology related to cloud computing has developed recently. In this regard, we confirmed that data transmission issues and security issues have emerged. Furthermore, we also found out that trends for video issues to recognize the image and smart city have emerged.

In the future, we will develop a system that can automatically perform trend analysis based on our method proposed in this study. The system will allow users to conduct trend analysis. Additionally, a research to predict the time to publish international standards for future technologies using machine learning will be implemented.

References

[1] W. Zhaohua, E. H. Norden, R. L. Steven and P. Chung-Kang, "On the trend, detrending, and variability of nonlinear and nonstationary time series", Proceedings of the National Academy of Sciences., vol. 38, **(2007)**, pp. 14889-14894.

[2] S. D. Grantham and L. H. Ungar, "Comparative analysis of qualitative models when the model changes", AlChE Journal., vol. 37, no. 6, **(1991)**, pp. 931-943.

[3] M. R. Maurya, R. Rengaswamy and V. Venkatasubramanian, "Fault Diagnosis by Qualitative Trend Analysis of The Principle Components", Journal of Chemical Engineering Research and Design., vol. 83, no. 9, **(2005)**, pp. 1122-1132.

[4] M. David, "Practical Strategies for Combining Qualitative and Quantitative Method: Applications to Health Research", Qualitative Health Research., vol. 8, no. 3, **(1998)**, pp. 362-376.

[5] D. Blei, "Probabilistic Topic Models", Communications of the ACM., vol. 55, **(2012)**, pp. 77-84.

[6] W. Xuerui and M. Andrew, "Topics over time: a non-Markov continuous-time model of topical trends", Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and data mining, Philadelphia, Pennsylvania, **(2006)**, August 20-23.

[7] J.H. Lau, N. Collier and T. Baldwin, "On-line Trend Analysis with Topic Models: #twitter trends detection topic model online", Proceedings of COLING, **(2012)**, pp. 1519-1534.

[8] ITU, "ITU-T Recommendations", Retrieved July 27, **(2017)**, from http://www.itu.int/en/ITU-T/publications/Pages/recs.aspx.

[9] Garter, "The Top 10 Strategic technology Trends for 2013", **(2013)**, from https://www.gartner.com/newsroom/id/2209615.

[10] Garter, "The Top 10 Strategic technology Trends for 2014", **(2014)**, from https://www.gartner.com/newsroom/id/2603623.

[11] Garter, "Gartner's Top 10 Strategic Technology Trends for 2015", **(2015)**, from https://www.gartner.com/smarterwithgartner/gartners-top-10-strategic-technology-trends-for-2015/.

[12] ITU, "ITU-T Recommendation X.1601: Security framework for cloud computing", Retrieved Oct, **(2016).**

[13] ITU, "ITU-T Recommendation Y.3500: Information technology – Cloud computing – Overview and Vocabulary", Retrieved Aug, **(2014)**.

[14] ITU, "ITU-T Recommendation Y.4115: Reference architecture for IoT device capability exposure", Retrieved Apr, **(2017)**.

Authors

Myeong-Ha Hwang, he received the B.S. degree in information and communication engineering from Chungnam National University, Daejeon, Korea, in 2015 and is currently pursuing the M.S. degree in information and communication network technology at University of Science and Technology (UST), Daejeon, Korea. Since 2016. He has been an UST Graduate Student with the Protocol Engineering Center (PEC), ETRI, Daejeon, Korea. His research interests include Text Mining, Cloud Computing and Big Data Analysis.

Dr. Suwook Ha, he has been working for Electronics and Telecommunications Research Institute (ETRI) since 2008 and working as a researcher in the field of ICT standardization. He specializes in software architecture including Geospatial Information System, Big Data, Cloud Computing, etc. Currently he is an editor of JTC 1 WG 9 and ITU-T SG13, a vice-chair of NGIS PG of TTA, and secretary of Big Data SPG of TTA. He is also working with Government to support the Next Generation Computing.

Minkyo In, he received the B.S. degree and M.S. degree in information and communication engineering from Chungnam National University, Daejeon, Korea. He has been working for Electronics and Telecommunications Research Institute (ETRI) since 2000. He has been working as a researcher in the field of ICT standardization (ITU-T SG13, ITU-T SG16, etc.). His research interests include Web of things, Cloud Computing and Big Data.

Dr. Kangchan Lee, he has been working for Electronics and Telecommunications Research Institute (ETRI) since 2001 and working as a professor in information and communication network technology at University of Science and Technology, Daejeon, Korea. His research interests are Web, Cloud Computing, Big Data, Blockchain/DLT, and Artificial Intelligence. Since 2005, he is working with ITU-T to develop the several editorships in Study Group 13 of ITU-T, he served vice chairman of ITU-T FG-Cloud. Also he is the Rapporteur of Q17 of Study Group 13 in ITU-T since 2010. Also, he has started as an Editor of ISO/IEC 19941(Cloud Computing – Interoperability and Portability) in JTC 1 SC 38 WG 4.