International Journal of Grid and Distributed Computing
Vol. 11, No. 5 (2018), pp.91-100
http://dx.doi.org/10.14257/ijgdc.2018.11.5.08

Related Document Extraction based on Topic Modeling using Cloud System

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

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

Abstract

Recently, studies on document recommendation and related document extraction using topic modeling have progressed. By using topic modeling, we can quickly understand the general overview of the documents. In this paper, we propose a method to extract related documents based on topic modeling using cloud system. The method consists of 8 steps. The proposed method is applied to extract related documents for each topic. Experiments are conducted using ITU-T Recommendations. As a result of the experiments, it is confirmed that the extracted documents are related to each series including a subject.

Keywords: *Related Document Extraction, Topic Modeling, Cloud System*

1. Introduction

Finding a document among the huge amount of documents is not easy. However, these difficulties are being solved with the development of topic modeling techniques. In particular, studies on document summarization and document recommendation using Latent Dirichlet Allocation (LDA) have been actively conducted. Extracting topics using the LDA can help people quickly and easily understand large amounts of documents. There is also a need for a system that can automatically perform topic modeling and related document extraction. It will help users summarize and understand the contents of documents more easily and quickly.

In this paper, we propose a related documents extraction method based on topic modeling using cloud system. 679 International Telecommunication Union Telecommunication Standardization Sector (ITU-T) Recommendations from 2013 to 2016 are used for the experiments. Each document is preprocessed and constructed into a table in database. Then, a representative topic from each document is constructed into another table in database. After joining two constructed tables, documents related to each keyword and each topic are extracted using the method in this study. To apply this method, a cloud system is designed and implemented.

2. Related Work

Researchers who developed the information retrieval system used topic modeling to provide items that meet needs. Topic modeling is a statistical model designed by natural language processing and machine learning, and involves finding topics in many documents. In general, a document consists of several words, and each word conveys specific information through a grouping of words. As shown in Figure 1, it can organize word clusters in one document and extract topics representing each word cluster.

Received (December 6, 2017), Review Result (February 6, 2018), Accepted (April 9, 2018)
* Corresponding Author

ISSN: 2005-4262 IJGDC

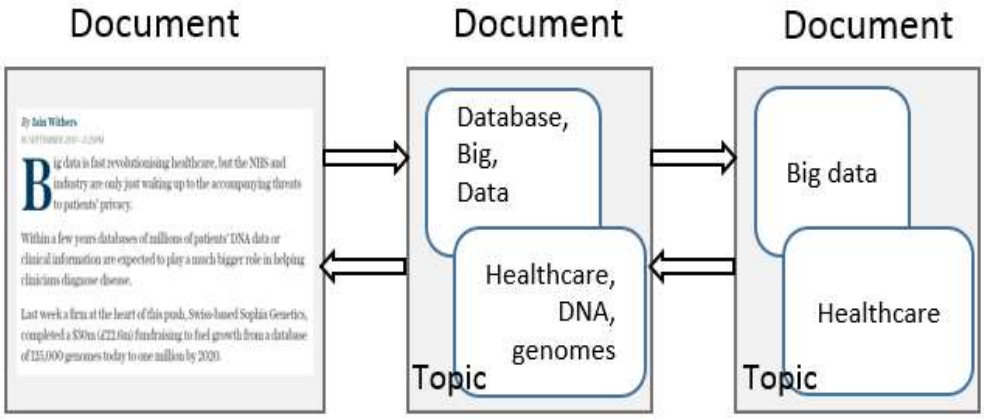

Figure 1. Document Representation of the Topic Modeling

Topic modeling is the process to generate document through each word. Probabilistic Latent Semantic Indexing (PLSI) is a typical algorithm among the many topic modeling algorithms[1]. PLSI is a mixed-decomposition-based. This estimates the coincidence rate in a mixture of conditional independent multinomial distribution. Each topic is defined as a mixture of V x M size. Therefore, kV + kM parameters are required for k subjects and increases linearly with M. The model tends to overfit when the size of the training set increases. LDA has been introduced to overcome the weaknesses of PLSI[2]. The LDA assumes that the subjects follow the dirichlet distribution. Therefore, there is an advantage that the overfit of model does not occur even if the size of the training set increases.

Since the introduction of LDA, researches have been conducted to summarize documents such as LDA-based multi-document summarization and latent dirichlet learning for document summarization[3][4]. Recently, it is confirmed that research related to document recommendation is proceeded such as LDA based integrated document recommendation model for e-learning systems, LDA-based personalized document recommendation, keyword extraction and clustering for document recommendation in conversations [5-7]. Furthermore, studies on system construction for document extraction and recommendation is proceeding such as a document recommendation system using a document-similarity ontology, personal document recommendation system based on data mining, content-based recommendation systems [8-10].

3. Related Document Extraction

3.1. Data Set

ITU-T Recommendation consists of a cover, summary and contents. The cover is composed of a title representing the document. The summary is made up of the content of the text. The contents contain descriptions of subject of the document. It consists of scope, reference, definitions, abbreviations, conventions, content, annex, bibliography, and is written in text, picture, *etc*.

In order to apply the method to extract related documents, we have collected 679 ITU-T Recommendations on Aug 29, 2017 from the ITU official website[11]. The ITU-T Recommendations consist 23 series from A series to Z series. Each series was published including each subject. However, we used 22 series from D series to Z series as shown in Table 1 because A series does not contain subject. ITU-T Recommendations from 2013 to 2016 was used.

Table 1. Series of ITU-T Recommendations

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

3.2. Design of Flow Chart

Flow Chart of the proposed document extraction method is composed of 8 steps as shown in Figure 2 and the description of each step is as follows.

1. Extract the representative topic

- Extract the topic in each document and normalize the word count in the topic

2. Topic Modeling

- Implement topic modeling after setting the scope of the documents and normalize the dirichlet parameter in the topic

3. Compare keywords results

- If the keywords from the results of step 1 and 2 are the same, send them to step 4, otherwise send them to step 5

4. Extract the weights of occurrence rate

- The weight are extracted by multiplying the occurrence rate of the same keywords

5. Compare documents results

- Select the topic for extraction of related documents (The selected topic contains 10 keywords)

- If the documents with the occurrence rate of each keyword are the same, send them to step 6, otherwise send them to step 7

6. Sum up the occurrence rate

7. Sort the cumulative occurrence rate in descending order

8. Extract related documents

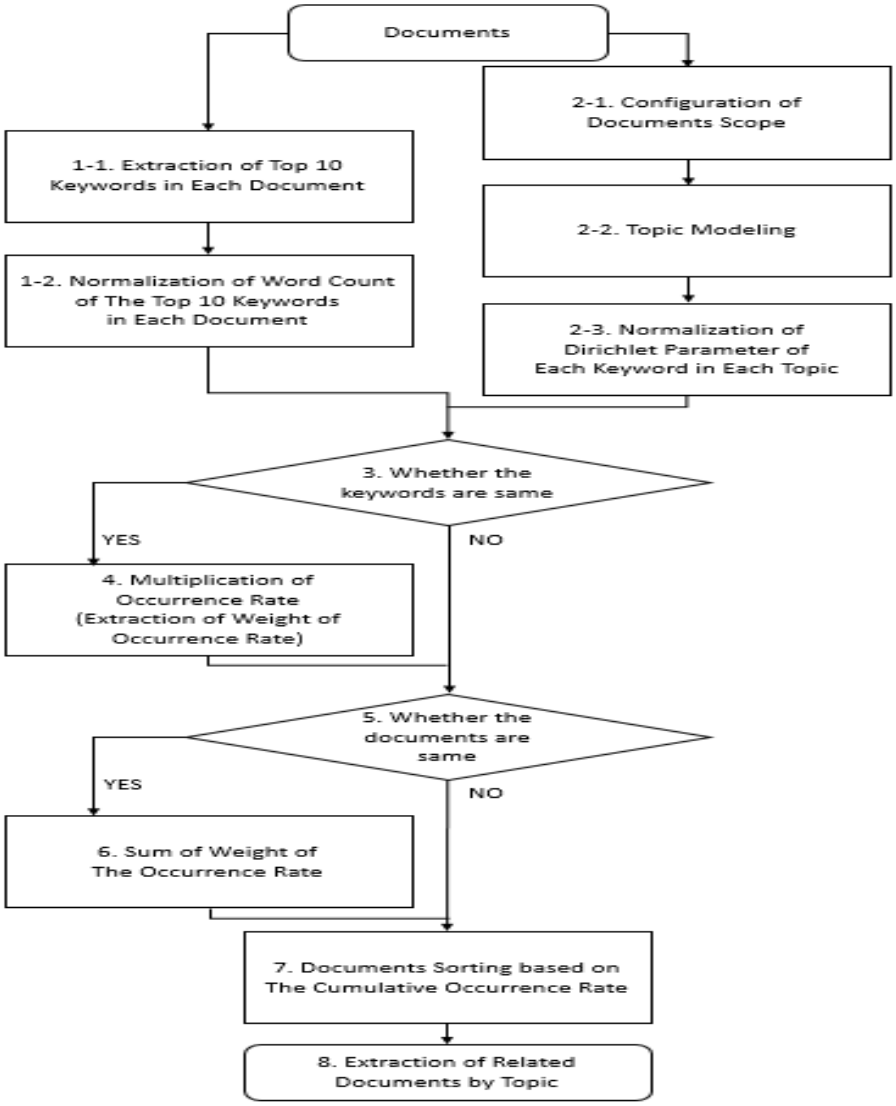

Figure 2. Flow Chart of Proposed Method

3.3. Design of Cloud System

We designed a cloud system to apply the method of related document extraction by topic. The overall design of the cloud system is shown in Figure 3. Topic modeling is performed after the scope of documents, the number of topics, and the number of iterations of the LDA have been set. The topic modeling results are constructed as a topic table. Then, the document table and the topic table are joined to extract related documents by keyword. If the keywords are the same, the occurrence rate is multiplied to extract the weight. The next step is to add up the occurrence rate if the documents from the keyword-related document extraction are the same. Finally, related documents for each topic are extracted after sort by descending order based on the total occurrence rate.

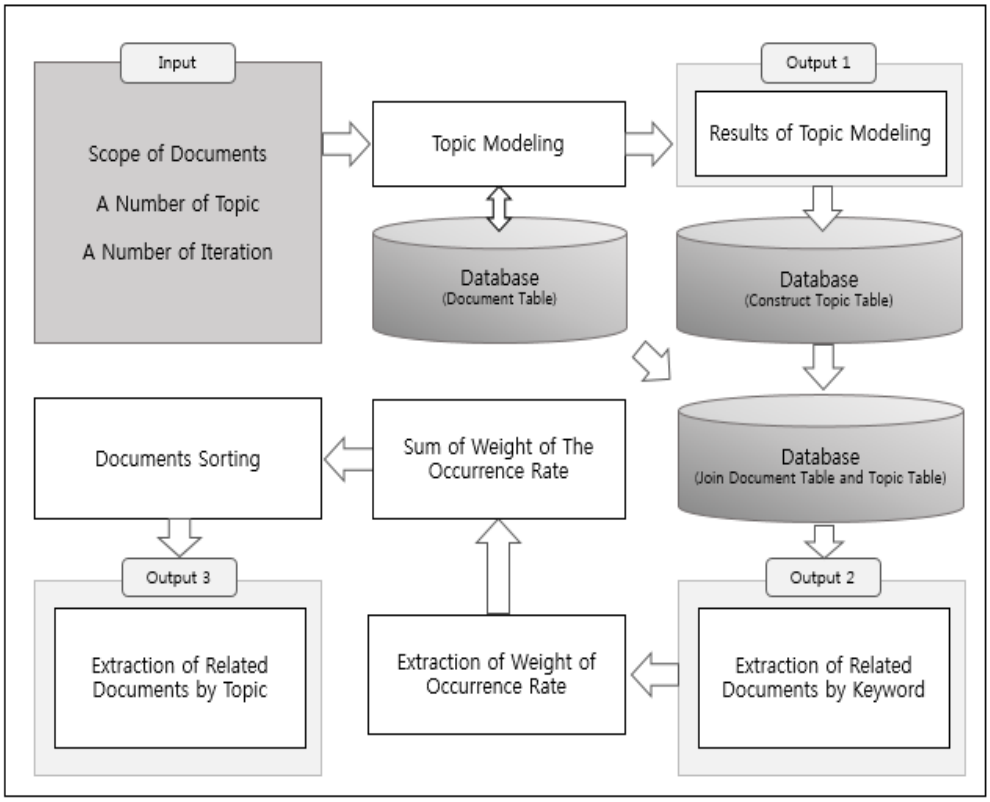

Figure 3. Design of Cloud System for Extracting Related Document by Topic

As shown in Figure 4, the document table is constructed. The name of each ITU-T Recommendation file is formatted as 'T-REC-E.1110-201301-I!!PDF-E.txt'. Therefore, 'E.1110' is extracted as Document ID, 'E' is extracted as Series, and '2016' is extracted as Year into table in database. The contents of the file are preprocessed. The preprocessing process is designed to construct the document table after three steps: Tokenization, Stop-word Removal, and Lemmatization.

The topic table is constructed by performing topic modeling on the document table. Representative topic which includes 10 keywords and word count of each keyword is extracted when topic modeling is performed. The number of word count of each keyword in the representative topic for each document is different because the size of each document is different. Therefore, occurrence rate is extracted so that all documents can be considered equally, and the equation is shown in (1) [12].

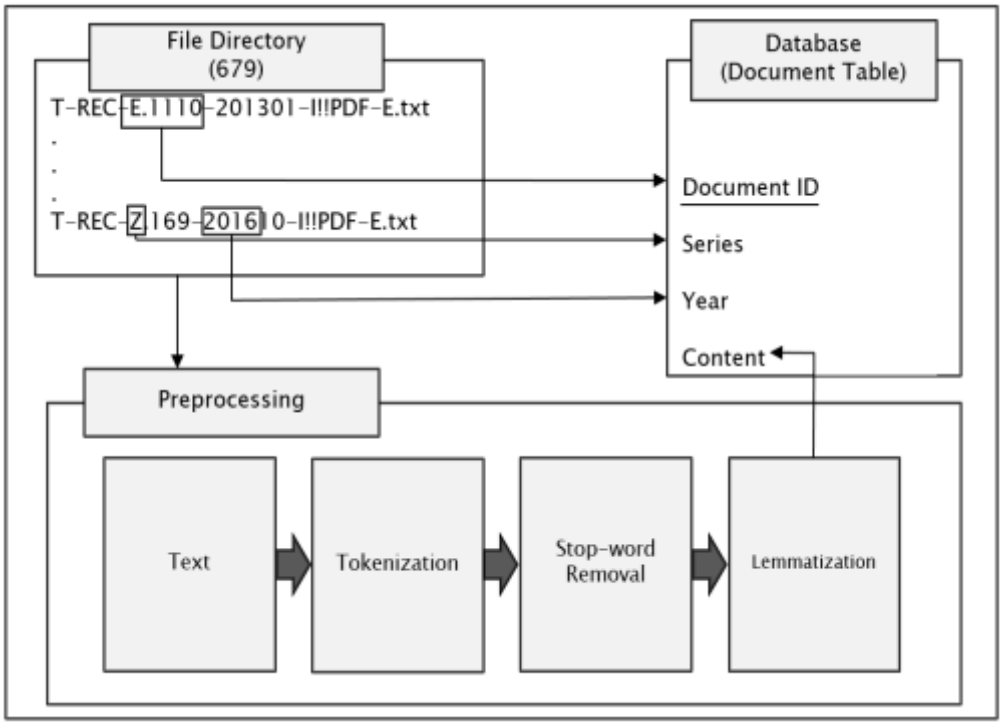

Figure 4. Configuration Process of Document Table

$$\text{Occurrence Rate}(\%) = \frac{Wordcount\ of\ each\ keyword_{\ n}}{\sum_{n=1}^{N} Wordcount\ of\ each\ keyword_{\ n}} * 100 \quad (1)$$

4. Experiments and Results

4.1. Results of Topic Modeling

The configuration for the topic modeling are as follows.

- Scope of the documents: 679 ITU-T Recommendations
- Period: 2013 – 2016
- The number of Topic: 10
- The number of Iteration of LDA: 1,000
- The number of Top Topic: 5

In order to experiments, a total of 679 ITU-T Recommendation are used from D series to Z series. The period is set from 2013 to 2016, which is the latest period specified by ITU-T. Ten topics are set for the experiment because the loglikelihood is highest when the number of topics is 10. Then, the number of iterations of LDA is experimented after the number of topics was set to 10. The experiment of iteration is proceeded to 100 units from 100 to 1500 times and is extracted by measuring the cosine similarity between the previous topic modeling result and next topic modeling result. As a result, the number of iterations is decided to be 1,000 because the degree of cosine similarity is more than 95%. The results of the topic modeling are shown in Table 2.

Table 2. Results of Topic Modeling

Period	Topic Name	Topic
2016 – 2013	Topic 1	Service, Cloud, Network, Information, Application, User, Management, Device, Security, Datum
	Topic 2	Energy, ICT, Datum, System, Network, Impact, Service, City, Information, Change
	Topic 3	TS, WWW, HTTP, Publish, ATIS, Service, TTC, 3G, Org, TTA
	Topic 4	Video, Quality, Element, IPTV, Content, Media, Model, Audio, Image, Event
	Topic 5	ONU, Stream, OLT, Message, PON, Bit, Frame, Upstream, Time, Channel

The documents related to topic 2 among the extracted topics, in which the slope of each year the sharpest increase, are extracted. The results are shown in Table 3. Topic 2 is relevant to the L series (Environment and ICTs, climate change, e-waste, energy efficiency, construction, installation and protection of cables and other elements of outside plant)[13-22].

Table 3. Results of Related Document Extraction by Topic

Topic	Rank	Document	Subject
Topic 2	1	L.1503	Use of information and communication technology for climate change adaptation in cities
	2	L.1501	Best practices on how countries can utilize ICTs to adapt to the effects of climate change
	3	L.1500	Framework for information and communication technologies and adaptation to the effects of climate change
	4	L.1410	Methodology for environmental life cycle assessments of information and communication technology goods, networks and services
	5	L.1430	Methodology for assessment of the environmental impact of information and communication technology greenhouse gas and energy projects
	6	L.1301	Minimum data set and communication interface requirements for data centre energy management
	7	L.1440	Methodology for environmental impact assessment of information and communication technologies at city level
	8	L.1330	Energy efficiency measurement and metrics for telecommunication networks
	9	Y.3035	Service universalization in future networks
	10	L.1601	Key performance indicators related to the use of information and communication technology in smart sustainable cities

5. Conclusion

In this paper, we proposed a method for extracting related documents based on topic modeling using cloud system. The proposed method is designed for table construction in database, topic modeling, related documents extraction by keyword, multiplication of occurrence rate in the case of the same keyword, and summing up the occurrence rate in the same documents. Then, the cloud system that can apply the proposed method for related document extraction is designed and developed. In order to experiments, ITU-T Recommendations was used. As a result of the experiment, 10 topics was extracted from international standards documents published from 2013 to 2016 and then top 5 topics was selected. Documents related to topic 2, in which the slope of each year showed the sharpest increase among extracted topics, was extracted. We confirmed extracted documents are related to L series.

The related document extraction method proposed in this study has limitations that are not personalized. In the future, we will study personalized recommendation method by complimenting the method of related documents extraction. Furthermore, we will develop a cloud system that can automatically implement personalized document recommendation methods. The system will be developed by using Deep Learning algorithm for personalized recommendation.

Acknowledgements

This work was supported by Institute for Information & Communication Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.R7116-16-1005, Development of Cloud Computing Interoperability Standards).

References

[1] T. Hofmann, "Probabilistic Latent Semantic Indexing", Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, California, USA, **(1999)**, August 15-19.
[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research., vol. 3, **(2003)**, pp. 993-1022.
[3] B. Ravindran and R. Arora, "Latent Dirichlet Allocation Based Multi-Document Summarization", Proceedings of the second workshop on Analytics for noisy unstructured text data, Singapore, **(2008)**, July 24-24.
[4] Y. L. Chang and J. T. Chien, "Latent Dirichlet Learning For Document Summarization", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, **(2009)** April 19-24.
[5] R. Nagori and G. Aghila, "LDA Based Integrated Document Recommendation Model for e-Learning Systems", Proceedings of International Conference on Emerging Trends in Networks and Computer Communications, Udaipur, India, **(2011)** April 22-24.
[6] T. M. Chang and W. F. Hsiao, "LDA-based Personalized Document Recommendation", Proceedings of Pacific Asia Conference on Information Systems, Jeju Island, Korea, **(2013)** June 18-22.
[7] H. Maryam and P. B. Andrei, "Keyword Extraction and Clustering for Document Recommendation in Conversations", IEEE/ACM Transactions on Audio, Speech, and Language Processing., vol. 23, **(2015)**, pp. 746-759.
[8] R. V. Nava, V. H. M. Dominguez and J. G. Montalvo, "A Document Recommendation System Using a Document-Similarity Ontology", IEEE Latin America Transactions., vol. 14, **(2016)**, pp. 3329-3334.
[9] S. M. Hsieh, S. J. Huang, C. C. Hsu and H. C. Chang, "Personal Document Recommendation System Based on Data Mining Techniques", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Beijing, China, **(2004)** September 20-24.
[10] M.J. Pazzani and D. Billsus, "Content-Based Recommendation Systems", The Adaptive Web: Methods and Strategies of Web Personalization., **(2007)**, pp. 325-341.
[11] ITU, "ITU-T Recommendations", Retrieved November 17, **(2017)**, from http://www.itu.int/en/ITU-T/publications/Pages/recs.aspx.
[12] M. H. Hwang, M. K. In, S. H. Ha and K. C. Lee, "TASIS: Trend Analysis System for International Standards", Proceedings of ITU Kaleidoscope, Nanjing, China, **(2017)** November 27-29.
[13] ITU, "ITU-T Recommendation L.1503: Use of information and communication technology for climate change adaptation in cities", Retrieved, **(2016)** June.

[14] ITU, "ITU-T Recommendation L.1501: Best practice on how countries can utilize ICTs to adapt to the effects of climate change", Retrieved, **(2014)** December.
[15] ITU, "ITU-T Recommendation L.1500: Framework for information and communication technologies and adaptation to the effects of climate change", Retrieved, **(2014)** June.
[16] ITU, "ITU-T Recommendation L.1410: Methodology for environmental life cycle assessments of information and communication technology goods, networks and services", Retrieved, **(2014)** December.
[17] ITU, "ITU-T Recommendation L.1430: Methodology for assessment of the environmental impact of information and communication technology greenhouse gas and energy projects", Retrieved, **(2013)** December.
[18] ITU, "ITU-T Recommendation L.1301: Minimum data set and communication interface requirements for data centre energy management", Retrieved, **(2015)** May.
[19] ITU, "ITU-T Recommendation L.1440: Methodology for environmental impact assessment of information and communication technologies at city level", Retrieved, **(2015)** October.
[20] ITU, "ITU-T Recommendation L.1330: Energy efficiency measurement and metrics for telecommunication networks", Retrieved, **(2015)** March.
[21] ITU, "ITU-T Recommendation Y.3035: Service universalization in future networks", Retrieved, **(2015)** June.
[22] ITU, "ITU-T Recommendation L.1601: Key performance indicators related to the use of information and communication technology in smart sustainable cities", Retrieved, **(2016)** June.

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.