International Journal of Advanced Science and Technology
Vol.116 (2018), pp.1-12
http://dx.doi.org/10.14257/ijast.2018.116.01

Entity-Driven Knowledge Analytics Platform for Science and Technology

Yuchul Jung[1], Jinyoung Kim[2*] Heung-Seon Oh[3], Dongjun Suh[4],
Jeong-Soo Kim[5], Seok-Hyong Lee[6], Kwang-Young Kim[7] and Jungsun Yoon[8]

[1]Department of Computer Engineering,
Kumoh National Institute of Technology, Gumi, Korea
[2,5,6,7,8]Korea Institute of Science and Technology Information, Korea
[3]School of Computer Science and Engineering,
Korea University of Technology and Education, Korea
[4]School of Convergence & Fusion System Engineering,
Kyungpook National University, Korea
[1]jyc@kumoh.ac.kr, [2]jykim@kisti.re.kr, [3]ohhs@koreatech.ac.kr,
[4]dongjunsuh@knu.ac.kr, {[5]mysticfact, [6]skyi, [7]glorykim, [8]jsyoon}@kisti.re.kr

Abstract

Research on the identification of persons, organizations, and locations in texts has been conducted in individual domains and application development cases, but there has been almost no research on platforms that enable the analysis of these data through interconnections. In this study, we developed functions to analyze the meta information of data and various contents and to automatically identify and analyze the identification information needed to link them. Furthermore, these functions were converted to open APIs, which were then mounted in an entity-drive knowledge analytics platform. To demonstrate the efficacy of the proposed platform, scenarios of various analytical application services are presented and the possibility of future expansions of the entity-driven analytic platform based on them is explained.

Keywords: *analytics platform, open APIs, knowledge sharing, data analysis, named entity*

1. Introduction

Since the early 2010s, a lot of efforts have been made to share data at home and abroad and the development and activation of shared platforms has become a major trend. Representative examples of shared platforms include CKAN [14] and Socrata [15] for public data sharing in the U.S. and the U.K. However, most platforms have focused on visualization and services rather than analytic functions. Recently, attempts have been made by MS Azure[16], Google Cloud[17], the Amazon Web Service (AWS)[18], etc. to provide analytics and development environ-ments that allow users to easily analyze data beyond data sharing.

In order to perform effective analytics of data and contents, it is critical to configure a meaningful knowledge network by discovering and interconnecting detailed properties and components. A representative research subject in this regard is linked open data. Reference [19] shows various datasets published in a linked data format. According to the recent trend of continuously expanding and diversifying information domains following the expansion of DBpedia in the past, more advanced incoming/outgoing links are being provided.

Received (January 4, 2018), Review Result (March 8, 2018), Accepted (March 12, 2018)

ISSN: 2005-4238 IJAST

However, shared contents include various types of data and have different properties. Thus, there are many limitations in closely interconnecting and utilizing them. For example, in order to connect with DBpedia at an effective knowledge level, everything must be prepared and mapped according to the ontology specifications [20] presented in DBpedia. Because newly created contents have a diversity of information and different depths of knowledge, a more realistic method is to connect them after automatically analyzing the surface knowledge. In this study, a knowledge sharing platform that allows for knowledge processing according to the user's intention is proposed to interconnect content around entities.

The proposed platform has the following three characteristics:

1) Connects the contents that contain data by converting them into knowledge at the registration stage.

2) Provides tools for identification and analysis of the meta information of major data types loaded by users.

3) A basis for additional applications is provided through an application scenario using loaded content and analysis tools.

2. Related Work

2.1. Entity Identification

Various studies on entity identification are drawing increasing attention, and academic challenges such as Entity Recognition and Disambiguation (ERD) [1] related to entity identification are also being organized. The entity identification process in ERD consists of two steps. The first step is spotting, which discovers entity candidates in text, and the second step is disambiguation, which determines the type of tagging that is appropriate for each entity in sentences. The entity identification research generally follows the process of ERD.

The knowledge analytics platform proposed in this study selects persons, organizations, and terms as major identification entities, and constructs, applies, and links identification data for science and technology content and expands knowledge by developing proprietary entity identification technology. For characters and terms, the supervised method of identification is mainly used. Especially for character identification, studies on various supervised methods [2–5] and unsupervised methods [7] have been conducted. In this study, the character identification result of a study on a supervised method based on a joint researcher network [6] was used. Research on organization identification is somewhat insufficient in comparison to persons and terms [8–11]. The proposed platform uses identification results by a method using an automatic identification algorithm based on an organization information database [12–13].

2.2. Open Data Platform

The representative platforms for sharing data owned by government and public agencies include CKAN [14] and Socrata [15]. CKAN was developed by the nonprofit organization Open Knowledge Foundation (OKF) and is being widely used in over 40 countries including the U.K., the U.S., and Canada. Besides basic functions, such specialized functions as visualization and API extraction have been developed by combining with other open sources such as Drupal [21]. As for CKAN, the proprietary functions of the platform can be used, and it is possible to provide separate services with the CKAN API only.

DSpace [22] is a shared platform that is widely used for sharing the data of researchers. DSpace is free software that was jointly developed by the Massachusetts Institute of

Technology and HP Labs. It is now being used at more than 1,000 universities, higher education institutions, cultural organizations, and research centers.

2.3. Data Analytics Platform

Data analytics functions can be classified into basic filtering and visualization functions for showing the target data and in-depth analysis functions including user-driven data processing, machine learning, and analytical web services. In the U.S., data visualization services are being provided using the Socrata API [15], which is very useful for visualization. In addition, data and content convergence services are being provided through mash-ups with Google Map and other services.

The MS Azure platform [46] provides a programming environment that allows for easy data handling and also provides distributed analysis services, large-scale storage, and machine learning infrastructures that facilitate data conversion, interactive data visualization tools, machine learning, and the use of big data. Furthermore, it offers a convenient development and service environment as a cloud that enables the direct connection of classification and prediction service (*e.g.*, emotional classification)-based data learning.

Scival [23], which is a research competence analysis solution based on research data, allows for the analysis of various research performance information by researcher, field, and university using the vast academic information of Elsevier. It also provides researcher search and recommendation functions.

3. Knowledge Analytics Platform

The knowledge analytics platform proposed in this study constructs identification information for many different entities included in the contents (including data) entered/loaded through a standardized process. In addition, various analytical open APIs are developed using the information about the identified entities. Thus, it allows users to create new knowledge through data analysis rather than simply using the data. The architecture of the knowledge analytics platform in Figure 1 shows the overall flow in terms of data supply, processing, analysis, and application creation.

The proposed knowledge analytics platform consists of five layers: the Data Management Layer for loading content (including data), Data Preprocessing Layer for processing loaded data (including text and images), Data Supply and Analysis Tools Layer for extracting entity-based analysis results, Open API Management Layer for managing various analysis APIs, and Service Convergence Layer for creating new types of analysis functions or services by combining data and analysis tools.

3.1. Data Management Layer

The Data Management Layer supports the later utilization of data through data loaded by users and related information entered in the database. Figure 2 shows the process of loading random data to the platform. The user fills in the meta information of the data to be loaded in the predefined form. The digital object identifier (DOI), person identifier, organization identifier, and term identifier are assigned to the filled form according to a standard process. The DOIs are unique digital identifiers assigned by DOI Korea using the Open API.

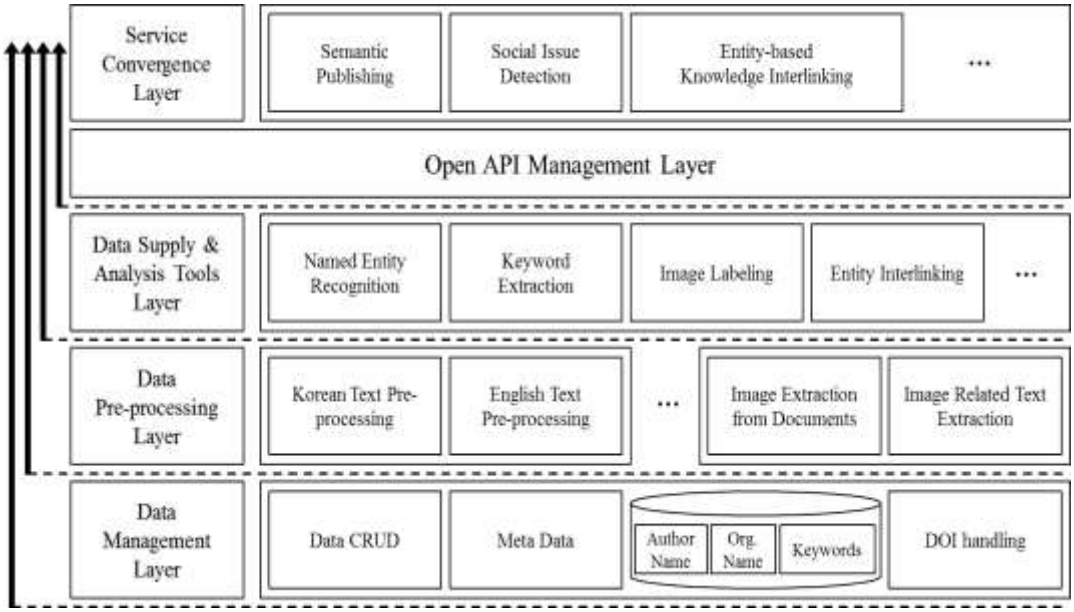

Figure 1. Architecture of Knowledge Analytics Platform

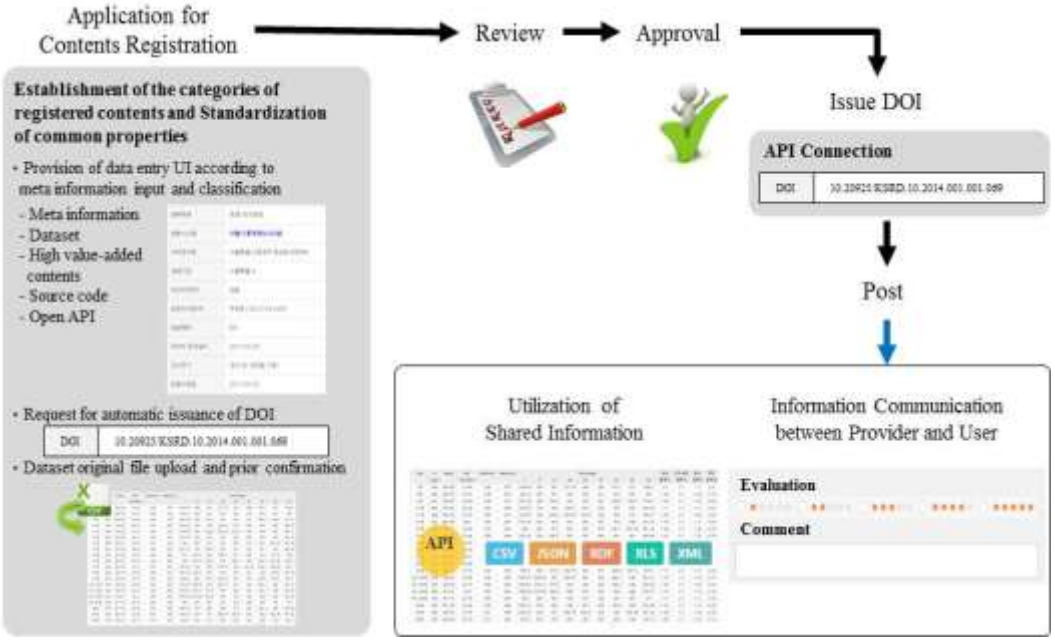

Figure 2. Content Registration Process

When the person/organization identifiers are entered using the identification information described in Section 3.1.1, multiple person/organization identifiers are recommended for the entered meta information. Then, the user determines the final information to be entered by manually selecting the accurate one among the recommended multiple candidates. The term identifiers are recommended using the automatic topic identification and keyword extraction API based on the identification information (Section 3.1.1) for keywords acquired from major science and technology content (papers, patents, and reports).

3.2. Basic DB Information for Entity Identification

The Korea Institute of Science and Technology Information (KISTI) has approximately 1,320,000 domestic papers in the science and technology fields, 210,000 science and technology research reports, 3,330,000 domestic patents, and 29,200,000 overseas papers.

KISTI has classified the major entities for identification regarding domestic science and technology content (papers, patents, and reports) by author, organization of the author, and main keywords (terms) summarizing individual content, and has constructed identification data by developing proprietary entity identification technology [6,12,13].

In the case of person identification, approximately 1,000,000 unique author names have been identified for papers, 430,000 for reports, and 1,900,000 for patents. As for organization identification, an organization information database for identification [12–13] has been built. It contains information for approximately 490,000 organizations including upper- and lower-level organizations in the hierarchy of organizations. Furthermore, 6,400,000 (around 85%) of 7,500,000 pieces of organization information of authors who have produced domestic science and technology content have been identified. In the case of term identification, approximately 720,000 keyword clusters have been constructed for approximately 740,000 items of domestic science and technology content that contain keywords.

The identification information for author names, organization names, and keywords are the meta information of the content that is most basically considered in the proposed knowledge platform and is the core basis for the interconnection of content.

3.3. Data Management Layer

The proposed knowledge analytics platform considers text and image data, and the Data Preprocessing Layer contains various modules for processing the data supplied to, loaded, and managed in the Data Management Layer. For text preprocessing, the Kokoma morphological analyzer [24] for Korea and the CoreNLP [25] of Standard University in the U.S. are mounted by default. These constitute the basis for analyzing the registered text content.

As for image data, preprocessing is only performed for PDF documents now, for which the open-source Apache PDFBox [26] is used. In particular, the data to be used in the API of the Analysis Tools Layer can be created using the features for separating text and images from the entered PDF documents and storing them.

3.4. Data Supply & Analysis Tools Layer

In the present study, the proposed knowledge analytics platform has converted various analysis functions to open APIs in accordance with the widely used RESTful API standard, and has mounted them in the platform. The proposed knowledge analytics platform considers the data APIs and text entity identification and analysis APIs listed in Table 1. The various APIs in this layer are serviced individually or in combination through the Open API Management Layer in Section 3.4 and the Service Convergence Layer in Section 3.5.

3.5. Open API Management Layer

The Open API Management Layer of the proposed knowledge analytics platform in Figure 3 performs management functions for various open API user authentication, key management for API calling, call management, and call log storing. The APIs listed in Table 1 have different characteristics for each implementation.

Table 1. List of Data and Analytics APIs Provided

Category	Name	Description
Data API	Person/organization/term identification data	A set of data to which unique identifiers have been assigned by controlling the researchers of papers, patents, and reports, their organizations, and the homographs and synonyms of research keywords
	Entity name tagging/identifier linking data	The persons, organizations, subjects, and area information in abstracts of academic papers and news articles in food sector are recognized, and data to which entity name tags and KISTI identifiers are assigned are provided.
	Learning data for entity name recognition system (food)	Learning data for automatic recognition of entity names in food sector
	Korean-Chinese-English dictionary	Data containing the Korean and English translations of Chinese science and technology terms
	Model-based relational data (network data)	Network data between entities produced through an analysis model
Text Entity Identification and Analysis API	SNEIK(Science & Technology Name Entity Identification Kernel) API	Automatic entity identification function for academic meta information such as papers, patents, and reports based on the person/organization/term identification data
	NERS(Named Entity Recognizer for Science and Technology Information) API	Deep learning-based automatic entity recognition function that identifies person, organization, and term entities that appear in the abstracts and news articles in science and technology sector
	NERS_Food(Deep learning-based named entity recognizer for food) API	Deep learning-based automatic entity recognition function that identifies food entities in science and technology news articles in the food sector
	ATiKA(Automatic Topic Key-words Annotation) API	Major keywords are selected and tagged considering the subject area in random input data (abstracts, explanations, etc.) (e.g., ICT, food, healthcare)
	EID (Event-based Issue Detector) API	Major issues are detected using the LDA topic analysis technique from the monthly news based on domestic news data collected for the last three years, and related entities are extracted around major events.
	NSCC(National Science & Technology Standard Classification System based Code Classifier) API	Top-k classification codes are recommended for random input data based on the national science and technology standard classifications (33 major categories, 371 middle categories)
Image Processing API	(Deep learning-based) Academic image entity recognition API	Objects in images are recognized and the entity names are extracted based on a model that has learned the images in academic papers by the deep learning technique.

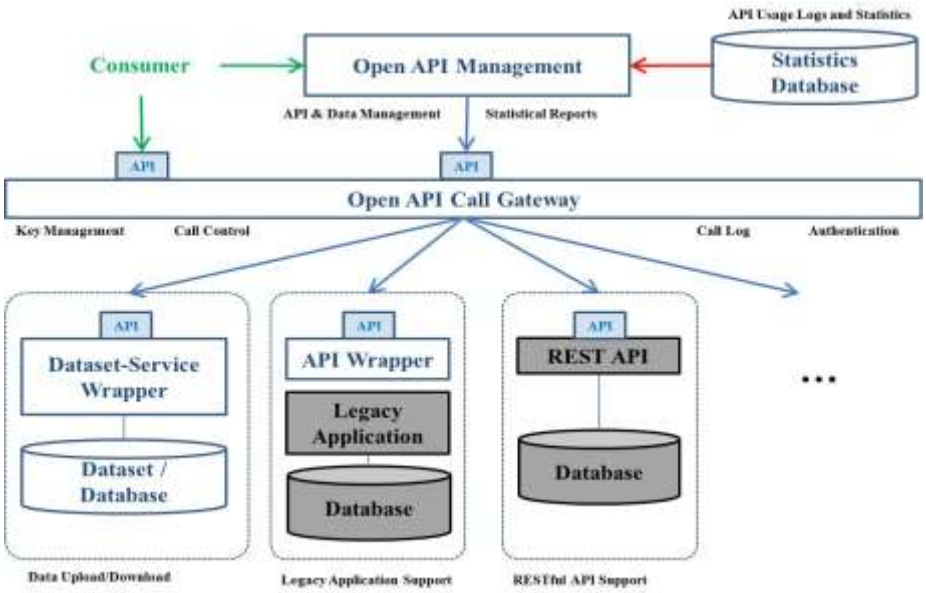

Figure 3. Open API Management Layer

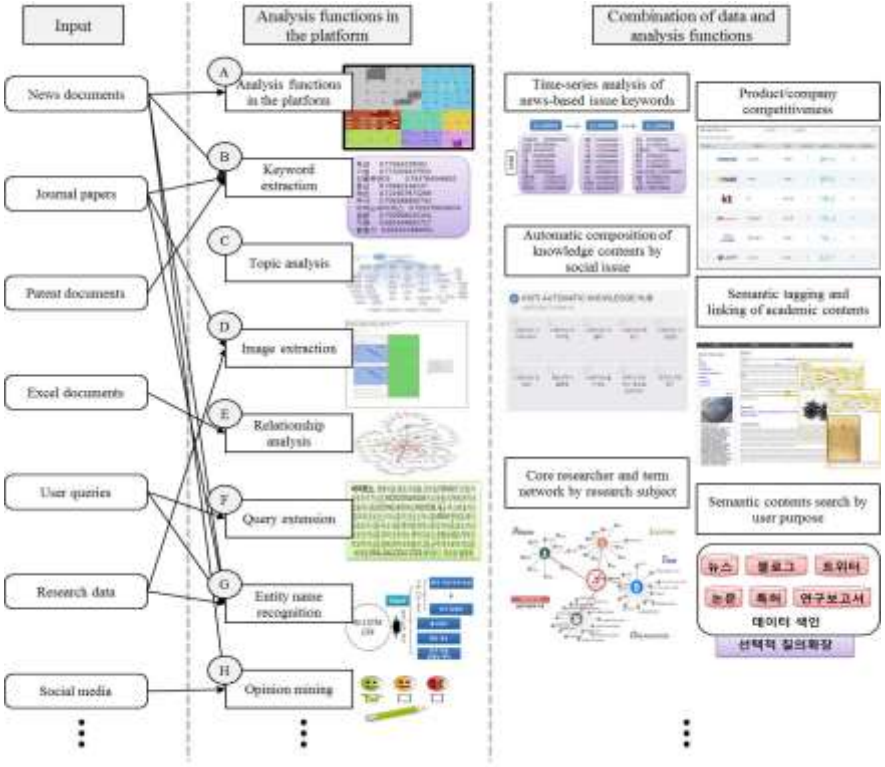

Figure 4. Service Composition According to the Combination of Data and Analytics API

Depending on the case, it may be a simple data download, and several seconds of standby time may be required for complex cases. APIs may be operated by implementing a wrapper in an existing application, or may be newly developed depending on the

RESTful API. Therefore, the Open API Management Layer creates an appropriate error code (including success and various failures) in consideration of the response time (the time elapsed from the calling until the result is returned to the user) depending on the call situation of each API.

3.6. Service Convergence Layer

Various analyses or services can be generated by applying the analysis function (Open API) provided by the knowledge analytics platform for input data. Figure 4 shows the flow of generating the required analysis or service by applying internal analysis functions for each input type. The primary output can be obtained by selectively applying the analysis function to input data, and another output can be generated by applying additional analysis functions as needed.

For example, monthly major social issues can be extracted using the news-event-based issue detection API from the news data, and can be displayed using the visualization API. Furthermore, major keyword extraction, image extraction, and entity name recognition can be selectively called for academic data, and applications for the output can be developed, including author keyword recommendation, paper data processing through paper image extraction, and connection of entity name recognition results with external knowledge base (*e.g.*, Wikipedia). In addition, the core researchers by research subject can be connected in terms of person-location-organization-term entities to support various entity-driven knowledge searches.

4. Discussion

4.1. Limitation in the Collection/Utilization of Diverse Data

Public and research data shared through other shared platforms are rarely used in a meaningful manner because they are different in purpose of use and storage format. This problem also occurs in the knowledge analytics platform proposed in this study. KISTI is continuously collecting and building science and technology content (papers, patents, and reports) serviced through the National Digital Science Library (NDSL), but the results of data collection efforts are insignificant with the exception of science and technology content. To promote the sharing of more diverse and wide-ranging data, strategic collaborations with external information procedures/providers (publishers, media, other public institutions, *etc.*,) are essential, and more consistent efforts should be made.

4.2. Need for Efforts to Add Various Analytical APIs

Currently, data provision (entity identification data, dictionary, synonyms, etc.) APIs, an API based on a person-organization-term-oriented relational model, text analysis API, text processing API, image processing API, and visualization API are included. However, most of these APIs have been implemented to process the data owned by KISTI and the collected news data. To activate the proposed knowledge analytics platform, a standardized API loading process is being established. In the future, the APIs of external developers can be mounted in the knowledge platform and serviced to a variety of users. Currently, the mounting of open APIs developed in Java and Python in the platform is supported.

4.3. Implementation of Robust Open API Management

More than 20 different open APIs were developed and tested. As a result, differences in API throughput occurred owing to various factors such as network condition, number of users, server-side capacity, and CPU performance. More research is needed on the

intelligent call control structure in the Open API Management Layer that detects the server-side load at the time of calling each open API.

4.4. Implementation of a Cloud-Based Flexible Service Combination and Running Environment

Most users can select and use appropriate analysis open APIs for input data. However, there are some limitations in building an environment for actual service using the analysis results. The MS Azure platform can easily show the functions developed by users as services in a cloud environment, but this feature is insufficient in this knowledge analytics platform.

5. Conclusion and Future Research

The present study proposed an analysis platform that is mounted with open APIs converted from analysis modules to enable major keyword extraction, subject analysis, and image extraction, as well as with the Science & Technology Name Entity Identification Kernel (SNEIK) based on the person, organization, and term identification data obtained from papers, patents, and reports. Currently, the functions of each analysis module have been stabilized and applied to the functional enhancement of existing services by connecting with the NDSL and NTIS services of the KISTI. In the future, the functions of this platform will be opened to outside users so that they can be used by various science and technology researchers.

Acknowledgments

This research was supported by Korea Institute of Science and Technology Information (KISTI).

This paper is a revised and expanded version of a paper entitled “Designing a Knowledge Analytics Platform for Science and Technology Contents” presented at The 1st International Conference on Convergent Research Theory and Technology, Jeju, Korea, Aug. 19-20.

References

[1] Y.-P. Chiu, Y.-S. Shih, Y.-Y. Lee, C.-C. Shao, M.-L. Cai, S.-L. Wei and H.-H. Chen, “NTUNLP approaches to recognizing and disambiguating entities in long and short text at the ERD challenge 2014”, Proceedings of the first international workshop on Entity recognition & disambiguation, Gold Coast, Australia, **(2014)** July 11.

[2] P. Kanani and A. McCallum, “Efficient strategies for improving partitioning-based author coreference by incorporating Web pages as graph nodes”, Proceedings of AAAI 2007 workshop on information integration on the Web, Vancouver, British Columbia, Canada, **(2007)** July 22-23.

[3] K.-H. Yang, J.-Y. Jiang, H.-M. Lee and J.-M. Ho, “Extracting citation relationships from web documents for author disambiguation”, Technical Report TR-IIS-06-017, **(2006)**.

[4] H. Han, G. Lee, H. Zha, C. Li and K. Tsioutsiouliklis, “Two supervised learning approaches for name disambiguation in author citations”, Proceedings of the 2004 joint ACM/IEEE conference on IEEE, Tucson USA, **(2004)** June 7-11.

[5] M. Yoshida, M. Ikeda, S. Ono, I. Sato and H. Nakagawa, “Person name disambiguation by bootstrapping”, Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland, **(2010)** July 19-23.

[6] J.-W. Seol, S.-H. Lee and K.-Y. Kim, “Author Disambiguation using Co-Author Network and Supervised Learning Approach in Scholarly Data”, Journal of Software Engineering and Its Applications, vol. 10, no. 4, **(2016)**, pp. 73-82.

[7] X. Yang, P. Jin and W. Xiang, “Exploring Word Similarity to Improve Chinese Personal Name Disambiguation”, Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Lyon, France, **(2011)** August 22-27.

[8] S.-H. Lee, “Study on the Construction of Identified Data of Author’s Affiliation in Academic Papers”, Journal of the Institute for Social Sciences, vol. 25, no. 4, **(2014)**, pp. 391-410.

[9] S.-H. Lee and S.-J. Kwak, "A Study on the Construction for Name Authority Data of the Korean Academic Papers", Journal of the Korean Biblia Society for Library and Information Science, vol. 21, no. 1, **(2010)**, pp. 105-118.
[10] S.-H. Lee and S.-J. Kwak, "Development and Evaluation of Authority Data based Academic Paper Retrieval System", Journal of the Society for Library and Information Science, vol. 46, no. 2, **(2012)**, pp. 133-156.
[11] E. Caron and H. Daniels, "Identification of Organization Name Variants in Large Databases using Rule-Based Scoring and Clustering with a Case Study on the Web of Science Database", Proceedings of the 18th International Conference on Enterprise Information Systems, Rome, Italy, **(2016)** April 25-28.
[12] J. Kim, S.-H. Lee, D. Suh, K.-Y. Kim and J. Yoon, "A Study on the Identification Algorithm for Organization's Name of Author of Korean Science & Technology Contents", Journal of Digital Contents Society, vol. 18, no. 2, **(2017)**, pp. 373-382.
[13] J. Kim, S.-H. Lee, D. Suh and Kwang-Young Kim, "A Study on the Method and System for Organization's Name Authorization of Korean Science and Technology Contents", Journal of Digital Contents Society, vol. 17, no. 6, **(2016)**, pp. 555-563.
[14] http://ckan.org, 28 June **(2017)**.
[15] https://socrata.com, 28June **(2017)**.
[16] https://azure.microsoft.com, 28June **(2017)**.
[17] https://cloud.google.com, 28 June **(2017)**.
[18] https://aws.amazon.com, 28 June **(2017)**.
[19] http://lod-cloud.net, 28 June **(2017)**.
[20] http://mappings.dbpedia.org/server/ontology/classes/, 28 June **(2017)**.
[21] https://www.drupal.org, 28 June **(2017)**.
[22] http://www.dspace.org, 28 June **(2017)**.
[23] https://www.elsevier.com/solutions/scival, 28 June **(2017)**.
[24] http://kkma.snu.ac.kr/documents/, 28 June **(2017)**.
[25] https://stanfordnlp.github.io/CoreNLP/, 28 June **(2017)**.
[26] https://pdfbox.apache.org/, 28 June **(2017)**.

Authors

Yuchul Jung, is an assistant professor in the Department of Computer Engineering at the Kumoh National Institute of Technology (KIT), South Korea. He received his Ph.D from the Department of Compuster Science in Korea Advanced Institute of Science and Technology (KAIST). His research includes natural langague processing, deep learning, and large-scale knowledge constsruction.

Jinyoung Kim, is a researcher at Korea Institute of Science and Technology Information, Daejeon, Korea. He is a Ph.D. student at Korea Advanced Institute of Science and Technology, Korea. He received a master's degree and a bachelor's degree from Sogang University, Korea. His research interests lies in entity identification, text data analysis, graph database, regular expression search, data mining and big data analysis.

Heung-Seon Oh, is an assistant professor in the School of Computer Science and Engineering at the Korea University of Technology and Education (KOREATECH), South Korea. He received his Ph.D from the Department of Compuster Science in Korea Advanced Institute of Science and Technology (KAIST). His research focus on developling artificiall intelligence techologies relevant to machine learnig, information retrieval, information extraction.

Dongjun Suh, is an assistant professor in the School of Convergence & Fusion System Engineering at the Kyungpook National University, South Korea. He received Ph.D. degree from Department of Civil and Environmental Engineering (Construction-ICT convergence) in Korea Advanced Institute of Science and Technology (KAIST). His current interest is smart control and system encompasses a variety of statistical techniques from machine learning, data mining and various theory that analyze current and historical facts to make predictions about future events

Jeong-Soo Kim, is a researcher at Korea Institute of Science and Technology Information, Daejeon, South Korea. He received his Ph.D. from the Department of Computer Science and Engineering in Pusan National University (PNU). His research focuses on developling artificiall intelligence techologies relevant to machine learning, information extraction, and natural language processing.

Seok-Hyong Lee, is a senior researcher at Korea Institute of Science and Technology Information, Daejeon, Korea. He received his Ph.D. from the Department of Library & Information Science in Chungnam National University, Korea. His current interest is information on processing, information analysis, bigdata analysis and entity identification.

Kwang-Young Kim, is a senior researcher at Korea Institute of Science and Technology Information, Daejeon, Korea. He received his Ph.D. from the Department of Library & Information Science in Chungnam National University, Korea. His current interest is personalized search system, information search and text mining.

Jungsun Yoon, is a principal researcher at Korea Institute of Science and Technology Information, Daejeon, Korea. She received her master's degree from the Department of Compuster Science in Korea Advanced Institute of Science and Technology (KAIST). Her current interest is information service, data analysis and Artificial Intelligence.

