J Inf Process Syst, Vol.15, No.5, pp.1096~1107, October 2019
https://doi.org/10.3745/JIPS.04.0137

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

# An Ontology-Based Labeling of Influential Topics Using Topic Network Analysis

Hyon Hee Kim* and Hey Young Rhee**

**Abstract**

In this paper, we present an ontology-based approach to labeling influential topics of scientific articles. First, to look for influential topics from scientific article, topic modeling is performed, and then social network analysis is applied to the selected topic models. Abstracts of research papers related to data mining published over the 20 years from 1995 to 2015 are collected and analyzed in this research. Second, to interpret and to explain selected influential topics, the UniDM ontology is constructed from Wikipedia and serves as concept hierarchies of topic models. Our experimental results show that the subjects of data management and queries are identified in the most interrelated topic among other topics, which is followed by that of recommender systems and text mining. Also, the subjects of recommender systems and context-aware systems belong to the most influential topic, and the subject of k-nearest neighbor classifier belongs to the closest topic to other topics. The proposed framework provides a general model for interpreting topics in topic models, which plays an important role in overcoming ambiguous and arbitrary interpretation of topics in topic modeling.

**Keywords**

Data Mining Ontology, Labeling of Topic Models, Ontology-based Interpretation of Topics, Topic Network Analysis

## 1. Introduction

Topic modeling [1] has attracted much attention for analyzing lots of textual documents. In particular, in the area of research trend analysis, topic modeling and its variations are mainly applied to scientific articles to find hot topics and cold topics. Blei [2] analyzed 17,000 journal articles by using latent Dirichlet allocation (LDA) model to find topics of "science" journals. Park and Song [3] found main research subjects and their upward trends and declines using LDA in library and information science. In this way, the topic modeling technique is effective in identifying trends of research topics according to time, but there are few studies considering relationships among research topics. In the interdisciplinary research, each topic is not independent, but strongly connected with each other. Therefore, finding relationships among topics is required.

In this research, to extract interrelationships among topics, social network analysis is applied to the network of topic models. Social network analysis is a well-known approach to analysis of relationships

Manuscript received May 29, 2018; accepted July 6, 2018.

**Corresponding Author:** Hyon Hee Kim (heekim@dongduk.ac.kr)
* Dept. of Statistics and Information Science, Dongduk Women's University, Seoul, Korea (heekim@dongduk.ac.kr)
**Dept. of Library and Information Science, Dongduk Women's University, Seoul, Korea (jonju@dongduk.ac.kr)

among entities [4]. In the case of text documents, keyword network analysis is generally used [5,6]. In the keyword network analysis, each keyword is used as node, and a network is constructed by connecting keywords which appear concurrently in the same document. Using the keyword network analysis, important keywords could be found, but influential topics consisting of several keywords could not be found.

In this paper, a topic network analysis proposed in our former research [7] is performed to find influential topics from scientific articles related to data mining research. After then, selected topics are interpreted based on the UniDM ontology which is developed for domain of data mining research. Usually, topic modeling has some difficulties in interpreting a topic based on keywords belonging to the topic. The same keywords have different meanings in different context, whereas different keywords might have the same meaning. In most of recent research on topic modeling, keywords belonging to a topic are used for understanding topics arbitrarily. Therefore, a general framework for interpreting the topics is also essential.

For this purpose, we proposed an ontology-based labeling framework. The UniDM ontology defines class hierarchies and relationships among classes based on Wikipedia category [8]. In addition, by text mining of Wikipedia articles, top important keywords are extracted and registered as instances of the defined classes. The UniDM ontology is used for understanding selected influential topics by mapping keywords into instances of classes in the UniDM ontology. A class with the most corresponding instances is interpreted by the subject of the topic.

In this research, we collected 2,103 abstracts of "data mining" related articles from top 5 computer journals from 1996 to 2015 to analyze subjects of data mining research. The selected journals are *IEEE Transactions on Knowledge and Data Engineering*, *Data Mining and Knowledge Discovery*, *Information Sciences*, *Very Large Data Bases*, and *Expert Systems with Applications*. Also, the UniDM ontology was built based on "data mining" articles in Wikipedia. The keywords belonging to the selected topic are mapped to the instances of the UniDM ontology and subjects are extracted from the classes based on the instances.

Our experimental results show that the topic with the highest value of degree centrality contains keywords related to recommender systems and text mining, and the topic with the highest value of betweenness centrality contains also recommender systems and context-aware systems. Finally, the topic with the highest value of closeness centrality contains K-nearest neighbors and context-aware systems.

The contribution of this paper is in the development and validation of the ontology-based labeling framework for interpreting influential topics resulted in the topic network analysis. First, by combining topic modeling and social network analysis, influential topics which are connected with other topics extracted, Second, based on the UniDM ontology, the selected topics are interpreted in a semi-automatic way, which makes the interpretation of those topics clear.

The remainder of this paper is organized as follows. In Section 2, we mention related work, and in Section 3, we explain the ontology-based labeling of topics in detail. In Section 4, we show experimental results and finally in Section 5, we give a concluding remarks.

## 2. Related Work

Topic modeling has attracted much attention as a useful tool for finding latent topics from document collections. There are two main concerns in the topic modeling research. One is interpretation and labeling of topics, and the other is finding influential topics from the topic models. In subsection 2.1, we explain ontology-based approaches to topic modeling, and in subsection 2.2, we present a various algorithm for automatic labeling of topics. Finally, in subsection 2.3, we mention correlation of topic models using social network analysis.

## 2.1 Ontology-based Approaches to Text Mining

Research on ontologies has been successfully applied to the research field of information retrieval and image retrieval [9,10]. Vijayarajan et al. [9] developed a framework consisting of an object-attribute-value extraction procedure from a natural English language query. Once a text is passed in the framework, it is broken down into clauses, and then analyzed providing characteristic properties, such as the part of speech, synonyms, hypernyms, and hyponyms. The ontology is used for both text retrieval and image retrieval. Also, an ontology-based approach to understanding the meaning of web interface signs is proposed [11]. The research showed that ontology mapping is important to user interface design and to understand semantics of interface signs.

Recently, research on classification of textual documents or clustering texts has used an ontology. Allahyari et al. [12] developed a domain ontology based on Wikipedia, and using the domain ontology, measured semantic similarities among documents for classification. Also, Fodeh et al. [13] showed that a predefined ontology improves performance of document clustering. In this point of view, the proposed ontology-based framework for understanding topics based on topic modeling is essential because the topics are represented by set of keywords. Wikipedia which is one of the world’s largest knowledge sources is widely used for developing ontologies in an automatic way or in a semi-automatic way [14]. In particular, Wikipedia’s category structure makes it easy to find conceptual hierarchies. In addition, infobox in Wikipedia defines conceptual relationships, and thus it can be used for class definition. In this research, by text mining of Wikipedia articles related to “data mining”, category structure, and infobox, classes and their instances are defined. OntoLDA is an ontology-based topic model for semantic tagging [15]. Wikipedia’s category network has been used for interpreting topic models. Their experimental results showed that the Wikipedia’s hierarchical ontology can be successfully used for semantic tagging of documents.

## 2.2 Automatic Labeling Techniques of Topic Models

Since probabilistic topic models are unsupervised analysis techniques of large document collections, interpreting a set of latent topics generally depends on human experts. Chang et al. [16] proposed a method for measuring the interpretability of a topic model using word intrusion and topic intrusion. While traditional metrics for validation of coherence and relevance of topic models are negatively correlated with the measures of topic quality, word intrusion and topic intrusion are able to measure evaluations depending on real-world task performance.

More recently, automatic labeling techniques of topic models have been researched [17-19]. Lau et al. [17] proposed a method for automatically labelling topics. A label candidate set is generated from the top-ranking topic terms, titles of Wikipedia articles containing the top-ranking topic terms, and sub-phrases extracted from the Wikipedia article titles. Magatti et al. [18] used probabilistic topic extraction (PTE) to analyze the content of documents and the meaning of words. Topic hierarchies are defined and topic labeling is performed on the topic hierarchy. Usually, topics have hierarchies, and thus automatic labeling of hierarchical topics is performed [19]. Using the structural relation among the topics, candidate labels are selected, and labels are ranked based on the topical structure.

## 2.3 Topic Models with Social Network Analysis

Although LDA is a useful tool for the statistical analysis of document collections, it is not able to extract interrelationships among topics. In [20], the authors proposed a correlated topic model (CTM), which extracts correlation among topics via the logistic normal distribution. The CTM gives better predictive performance and uncovers interesting descriptive statistics for facilitating browsing and search.

Research on social network analysis using topic models has been done [21,22]. Cha and Cho [21] applied LDA to the relationship graph in a social network analysis. Their experimental results showed that the benefits of applying LDA to clustering applications. Also, topic-based social network analysis is performed on the dark web [22]. Topic-based network is constructed, and topics are extracted from the topic-based social network. Their approach showed that social aspects of topic network is able to measured. Also, network regularization was done on topic modeling [23]. Based on the network regularization of topic models, author-topic network was analyzed and community detection was performed the author-topic network. Their experimental results showed that the proposed approach is effective to extract topics, discover communities and network structure.

# 3. The Ontology-Based Labeling of Topics

In this section, we describe the proposed ontology-based labeling technique for topic models. In subsection 3.1, we give an overview of the framework, and in subsection 3.2, we explain the UniDM ontology in detail. Finally, in subsection 3.3, we describe social network analysis of topics via topic modeling.

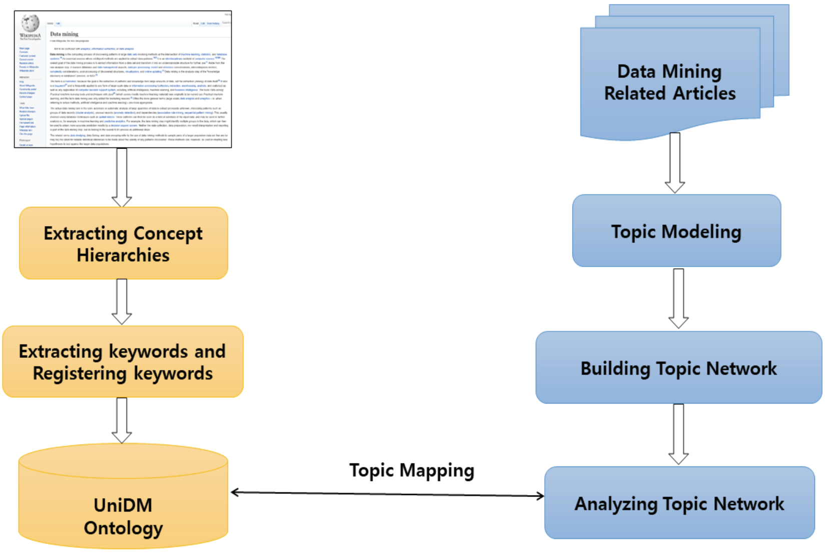
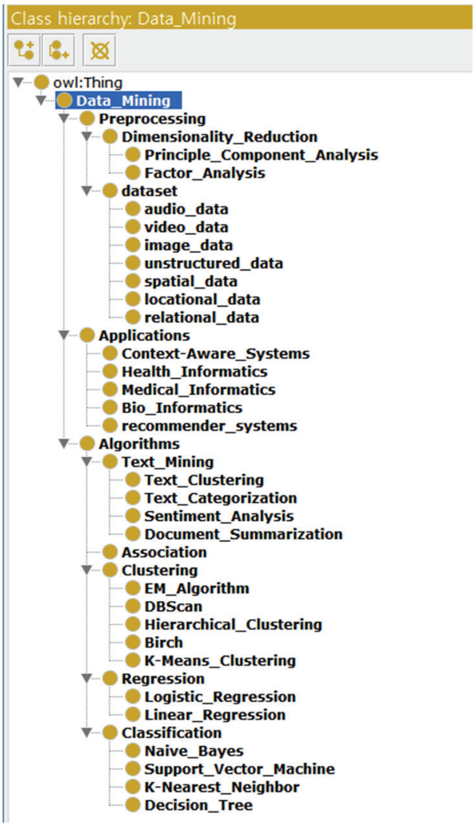


**Fig. 1.** Overview of the framework.

## 3.1 Overview of the Framework

Fig. 1 shows that the proposed ontology-based framework for understanding influential topics based on topic network analysis which is a social network analysis of topics in topic modeling. First, scientific articles related to "data mining" are collected from top 5 journals registered in Web of Science (http://isiknowledge.com), and abstracts of the articles are extracted. Second, nouns from abstracts are extracted from the abstracts, and then LDA topic modeling is performed. Third, social network analysis is performed on the topics from topic modeling. If topics have the same keywords, then the topics are connected with a weight value calculated by the number of common keywords. Once a social network of topics called topic network in this research is constructed, social network analysis is performed. As a result of the topic network analysis, influential topics are selected, and the selected topics are interpreted based on the UniDM ontology.

From the Wikipedia article about "data mining", concept hierarchies of data mining are extracted. Categories and infobox in Wikipedia are also considered in a semi-automatic way. After then important keywords with TF-IDF value are also extracted and are registered as instances. Correlation between keywords ant the target concepts is calculated and the keywords with over 0.5 correlation value are selected. The instances are used to determine an influential topic's class by keyword matching. Keywords in each topic are mapped to the UniDM ontology, and classes with matching keywords are used for labeling of the topics.



Class hierarchy: Data_Mining

- owl:Thing
  - Data_Mining
    - Preprocessing
      - Dimensionality_Reduction
        - Principle_Component_Analysis
        - Factor_Analysis
      - dataset
        - audio_data
        - video_data
        - image_data
        - unstructured_data
        - spatial_data
        - locational_data
        - relational_data
    - Applications
      - Context-Aware_Systems
      - Health_Informatics
      - Medical_Informatics
      - Bio_Informatics
      - recommender_systems
    - Algorithms
      - Text_Mining
        - Text_Clustering
        - Text_Categorization
        - Sentiment_Analysis
        - Document_Summarization
      - Association
      - Clustering
        - EM_Algorithm
        - DBScan
        - Hierarchical_Clustering
        - Birch
        - K-Means_Clustering
      - Regression
        - Logistic_Regression
        - Linear_Regression
      - Classification
        - Naive_Bayes
        - Support_Vector_Machine
        - K-Nearest_Neighbor
        - Decision_Tree

**Fig. 2.** Class Hierarchy in UniDM ontology.

## 3.2 The UniDM Ontology

To define the conceptual hierarchy of data mining research, Wikipedia which provides concept hierarchies and their meaning using categories and infobox is used. Fig. 2 shows the class hierarchy in UniDM using protégé ontology editor (http://protege.stanford.edu). The *Data Mining* class has three subclasses, i.e., *preprocessing* class, *application* class, and *algorithms* class. *The preprocessing* class defines data mining process of preprocessing and has two subclasses such as *dimensionality reduction* class and *data set* class. The application class defines application area of data mining, and has five subclasses, i.e., *context-aware system* class, *health informatics* class, *medical informatics* class, *bio informatics* class, and *recommender systems* class. The algorithms class defines data mining algorithms and has five subclasses, i.e., *text mining class*, *association* class, *clustering* class, *regression* class, and *classification* class. Each subclass might have further subclasses shown in Fig. 2.

To extract keywords which will be registered to the defined classes, keywords closely relate to the target class are selected. For this purpose, correlation between each keyword and the target class is calculated and keywords with over 0.5 correlation value are selected. The instances are used to determine an influential topic's class by keyword matching.

**Table 1.** Thirty topics using latent Dirichlet allocation

| | Keywords |
|---|---|
| Topic 1 | web, records, content, positive, pages |
| Topic 2 | procedure, review, papers, chain, reference |
| Topic 3 | scheme, hierarchical, variable, sparse, partial |
| Topic 4 | similarity, distance, index, metric, matching |
| Topic 5 | temporal, incremental, community, weighted, interactive |
| Topic 6 | social, matrix, stage, traffic, people |
| Topic 7 | hybrid, genetic, activities, Bayesian, recognition |
| Topic 8 | streams, regression, stream, credit, risk |
| Topic 9 | product, utility, mobile, recommendation, location |
| Topic 10 | detection, outlier, anomaly, fraud, attacks |
| Topic 11 | software, usage, construction, source, programming |
| Topic 12 | text, document, semantic, sentiment, weights |
| Topic 13 | sample, random, sampling, association, induction |
| Topic 14 | group, structure, gene, expression, biological |
| Topic 15 | attribute, rough, reduction, numerical, relative |
| Topic 16 | classifiers, ensemble, factors, map, criteria |
| Topic 17 | frequent, itemsets, memory, transaction, item |
| Topic 18 | event, relations, categorical, monitoring, log |
| Topic 19 | stock, financial, changes, market, forecasting |
| Topic 20 | item, approximation, generalize, constraint, minimal |
| Topic 21 | graph, relational, uncertain, ranking, graphs |
| Topic 22 | fuzzy, distributed, parallel, quantitative, architecture |
| Topic 23 | local, privacy, global, noise, control |
| Topic 24 | sequential, sequence, constraints, multidimensional, missing |
| Topic 25 | query, trees, queries, kmeans, gain |
| Topic 26 | feature, phase, svm, unsupervised, relevance |
| Topic 27 | clusters, objects, image, relation, regions |
| Topic 28 | series, spatial, points, neighborhood, sensor |
| Topic 29 | medical, sources, expert, instance, confidence |
| Topic 30 | customer, marketing, company, segmentation, churn |

## 3.3 Topic Network Analysis

In order to construct a topic network, topic modeling is performed. Table 1 shows 30 topics using LDA algorithm. To choose an appropriate number of topics, 10, 20, 30, and 40 number of topics are tested respectively and among them 30 topics are chosen because topic's meaning is clearly separated. Based on the 30 topics, topic network analysis is performed.

Algorithm 1 shows the TNA algorithm which is modified from the TNC algorithm [7] presented our former study. In the topic network, a topic serves as a network node, and the number of keywords appearing in other topics concurrently serves as a weight of the edge connecting nodes.

**Algorithm 1.** Algorithm for Topic Network Analysis

```
Input: a set of abstracts A
Output: topicDCentrality, topicBCentrality, topicCCentrality
begin initialization
    DM[$f_{i,j}$]          null;  //for document-term matrix
    DW[$w_{i,j}$]          null; //for document-term matrix with TF-IDF value
    M[i,j]                 null; //for LDA models
    W[$w_{i,j}$]           null; //for Topic Weights matrix
    abstractRawData, abstractCorpus, abstractCorpusClean      null;
    topicDCentrality, topicBCentrality, topicCCentrality       null;
end
begin preprocessing
  abstractRawData <- Read A;
  abstractCorpus <- Convert abstractRawData into abstractCorpus;
    while not untill end of line of the abstractCorpus do
        Remove special character;
        Remove numbers;
        Remove stopwords;
    end
  abstractCorpusClean <- abstractCorpus
end
begin create Topic Network
  abstractNouns <- extrat nouns from abstractCorpusClean
  DM[$f_{i,j}$] <- generate document-term matrix from abstractNouns
  DW[$f_{i,j}$] <- $f_{i,j}$ * log ( n / $df_{i,j}$ ) >= k
  M[i,j] <- create LDA model
    for each i
        for each j
              common <- intersect M[,i] with M[,j]
              if length(common) >= 1 then
                  T[$t_{i,j}$] <- length(common)
              else T[$t_{i,j}$] <- 0
        endfor
      endfor
begin perform Topic Network Analysis
  topicDCentrality <- assign the topic with the highest degree centrality
  topicBCenrality <- assign the topic with the highest betweenness centrality
  topicCCentrality <- assign the topic with the highest closeness centrality
end
```

First, abstracts of papers are preprocessed (line 7–16). From the corpus of abstracts, nouns are extracted and document-term matrix is generated. After then, document-term matrix is modified with the TF-IDF weight. Based on the document-term matrix, LDA is performed. A topic network is constructed with common words as weighted links (line 17–29). Social network analysis is performed on the topic network and the basic three measurements [4], i.e., degree centrality, betweenness centrality, and closeness centrality, are calculated on each topic. Topics with the highest degree centrality, betweenness centrality, and closeness centrality are selected.

Fig. 3 shows the topic network graph using the TNC algorithm with 30 topics. The degree centrality measures the number of connected nodes, and thus the topic with highest degree centrality can be considered as the most important topic. The betweenness centrality measures mediating each node, and the topic with the highest betweenness centrality can be considered as an important interdisciplinary research. The closeness centrality measures distances with all other nodes, and the topic with the highest closeness centrality can be considered as a closely related topic with other topics. As a result of topic network analysis, topic 25 (T25 in Fig. 3) has the highest degree centrality value, while topic 9 (T9 in Fig. 3) has the highest betweenness centrality value. Finally, topic 28 (T28 in Fig. 3) has the highest closeness centrality value.

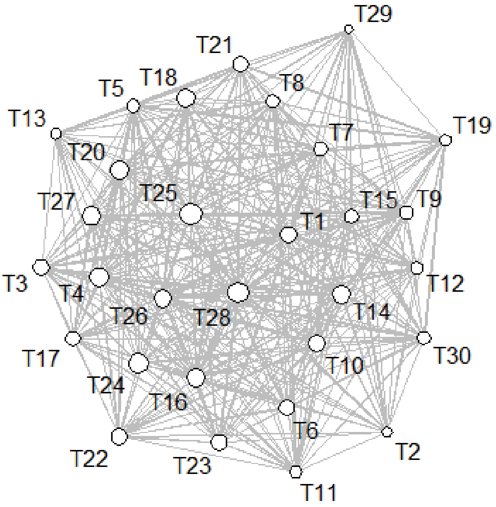


**Fig. 3.** Topic network graph.

## 4. Experimental Results

To prepare data for this research, we collected 2,103 abstracts of "data mining" related articles from top 5 computer journals published from 1996 to 2015. To find research subjects of the topics represented by keywords, each keyword is mapped into instances of classes in the UniDM ontology. Table 2 shows top 30 keywords belonging to topic 25, topic 9, and topic 28, respectively. Topic 25 is the topic with the highest value of degree centrality and topic 9 is the topic with the highest value of betweenness centrality. Finally, Topic 28 is the topic with the highest value of the closeness centrality.

**Table 2.** Top 30 keywords in Topic 25, 9, 28

| | Keywords |
|---|---|
| T25 | query, trees, k-means, gain, queries, summarization, optimized, proposal, language, extension, special, response, simulated, indexing, summarize, organization, guide, commercial, rank, filter, exhibit, compress, summary, ordered, access, relational, differences, operation, fixed, searching |
| T9 | product, utility, mobile, location, recommendation, recommendations, products, filtering, recommender, collaborative, personalized, preferences, e-commerce, personal, service, real-time, profits, guide, behaviors, sale, stores, digital, wireless, similarities, life, experience, taxonomy, rapidly, rank, producing |
| T28 | series, spatial, points, neighborhood, representations, neighbor, sensor, nearest, valid, raw, short, subsets, segment, simulated, composed, close, Euclidean, difference, wireless, discrete, indexing, analytical, location, searching, leading, evidence, filter, compares, per, incrementally |

First, let us look at the research subjects of topic 25 with the highest value of degree centrality in detail. Table 3 shows research subjects based on the keyword mapping between keywords in topic 25 and instances in the UniDM classes. The class with the most matching keywords is *data set* class, followed by *recommender systems* class and *document summarization* class. The *decision tree* class and *k-means* class have also matching keywords. Therefore, we conclude that data set is the most connected node with other nodes, which means that lots of research handled data set rather than research topic.

**Table 3.** Research subject in Topic 25

| Class | Matching keywords |
|---|---|
| Data set | query, queries, relational, ordered, optimized, indexing, access, differences, operation, searching, fixed, language, extension |
| Recommender systems | organization, commercial, filter, response |
| Document summarization | summarization, summarize, summary, rank |
| Decision tree | tree, gain |
| K-means | k-means |
| N/A | proposal, special, simulated, guide, exhibit, compares |

**Table 4.** Research subject in Topic 9

| Class | Matching keywords |
|---|---|
| Recommender systems | product, recommendation, rank, products, service, filtering, sales, similarities, recommendations, filtering, recommender, behaviors, collaborative, personalized, service, preferences, e-commerce, personal |
| Context-aware systems | mobile, location, real-time, wireless, life, experience |
| Text mining | taxonomy |
| N/A | utility, profits, guide, digital, stores, rapidly |

Next, consider Topic 8 which has the highest value of betweenness centrality. According to the keyword matching shown in Table 4, *recommender system* class has overwhelmingly matching keywords, followed by *context-aware systems* class. It is concluded that applications of data mining mainly mediate other research subjects.

Finally, let us look at topic 28, which has the highest value of closeness centrality shown in Table 5. The data set class has the most matching keywords, followed by *k-nearest neighbor* algorithm class and then *context-aware systems* class.

**Table 5.** Research subject in Topic 28

| Class | Matching keywords |
|---|---|
| Data set | Representations raw, subsets, valid indexing, searching, short, discrete |
| k-nearest neighbor | Neighborhood, neighbor, nearest, close, Euclidean, segment, difference |
| Context-aware systems | Series, spatial, points, wireless, location, sensor |
| N/A | simulated, closed, composed, per, compare, incrementally, evidence, leading, analytical |

## 5. Conclusions

In this paper, we analyzed subjects of data mining research by combining topic modeling and social network analysis. Scientific articles related to data mining are collected and topic modeling is performed on the abstracts of the articles. Based on the topics and common keywords among topics, a topic network is constructed, and social network analysis is applied to the topic network.

As a result of the topic network analysis, three influential topics, i.e., a topic with the highest value of degree centrality, a topic with the highest value of betweenness centrality, and a topic with the highest value of closeness centrality are figured out. To interpret the three topic logically, the UniDM ontology is developed based on the data mining articles in Wikipedia, keywords in the topic are mapped to instances of the classes in the UniDM ontology.

Our experimental results show that recommender systems and document summarization are the most connected topics with other topics. Also, recommender systems are the topic which mostly mediates other topics. Finally, k-nearest neighbor algorithm is the closest topic to other topics. Our approach overcomes the limitation of topic modeling which cannot consider relationships among topics, and suggested a new paradigm for interpretation of results in topic modeling. The proposed ontology-based framework for interpreting topic models is expected to play an important role in interpreting results of big data analytics.

## Acknowledgement

This work was supported by Dongduk Women's University Research Grant.

## References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77-84, 2012.

[3] J. H. Park and M. Song, "A study on the research trends in library & information science in Korea using topic modeling," *Journal of the Korean Society for Information Management*, vol. 30, no. 1, pp. 7-32, 2013.

[4] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. New York, NY: Cambridge University Press, 1994.

[5] A. Duvvuru, S. Kamarthi, and S. Sultornsanee, “Undercovering research trends: network analysis of keywords in scholarly articles,” in *Proceedings of 2012 9th International Conference on Computer Science and Software Engineering (JCSSE)*, Bangkok, Thailand, 2012, pp. 265-270.

[6] H. H. Kim, D. Kim, and J. Jo, “Patent data analysis using clique analysis in a keyword network,” *Journal of the Korean Data and Information Science Society*, vol. 27, no. 5, pp. 1273-1284, 2016.

[7] H. H. Kim and H. Y. Rhee, “Trend analysis of data mining research using topic network analysis,” *Journal of the Korea Society of Computer and Information*, vol. 21, no. 5, pp. 141-148, 2016.

[8] Wikipedia [Online]. Available: http://www.wikipedia.org/.

[9] V. Vijayarajan, M. Dinakaran, P. Tejaswin, and M. Lohani, “A generic framework for ontology-based information retrieval and image retrieval in web data,” *Human-centric Computing and Information Sciences*, vol. 6, article no. 18, 2016.

[10] M. Lee, Y. S. Park, and J. W. Lee, “Image-centric integrated data model of medical information by diseases: two case studies for AMI and ischemic stroke,” *Journal of Information Processing Systems*, vol. 12, no. 4, pp. 741-753, 2016.

[11] M. N. Islam and A. N. Islam, “Ontology mapping and semantics of web interface signs,” *Human-centric Computing and Information Sciences*, vol. 6, article no. 20, 2016.

[12] M. Allahyari, K. J. Kochut, and M. Janik, “Ontology-based text classification into dynamically defined topics,” in *Proceedings of 2014 IEEE International Conference on Semantic Computing*, Newport Beach, CA, 2014, pp. 273-278.

[13] S. Fodeh, B. Punch, and P. N. Tan, “On ontology-driven document clustering using core semantic features,” *Knowledge and Information Systems*, vol. 28, no. 2, pp. 395-421, 2011.

[14] F. Wu and D. S. Weld, “Automatically refining the Wikipedia infobox ontology,” in *Proceedings of the 17th international conference on World Wide Web*, Beijing, China, 2008, pp. 635-644.

[15] M. Allahyari and K. Kochut, “Semantic tagging using topic models exploiting Wikipedia category network,” in *Proceedings of 2016 IEEE 10th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, 2016, pp. 63-70.

[16] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, “Reading tea leaves: how humans interpret topic models,” *Advances in Neural Information Processing Systems*, vol. 22, pp. 288-296, 2009.

[17] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, “Automatic labelling of topic models,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Portland, OR, 2011, pp. 1536-1545.

[18] D. Magatti, S. Calegari, D. Ciucci, and F. Stella, “Automatic labeling of topics,” in *Proceedings of 2009 9th International Conference on Intelligent Systems Design and Applications*, Pisa, Italy, 2009, pp. 1227-1232.

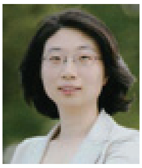
[19] X. L. Mao, Z. Y. Ming, Z. J. Zha, T. S. Chua, H. Yan, and X. Li, “Automatic labeling hierarchical topics,” in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, HI, 2012, pp. 2383-2386.

[20] D. M. Blei and J. D. Lafferty, “A correlated topic model of science,” *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 17-35, 2007.

[21] Y. Cha and J. Cho, “Social-network analysis using topic models,” in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, OR, 2012, pp. 565-574.

[22] G. L'huillier, H. Alvarez, S. A. Rios, and F. Aguilera, “Topic-based social network analysis for virtual communities of interests in the dark web,” *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 66-73, 2011.

[23] Q. Mei, D. Cai, D. Zhang, and C. Zhai, “Topic modeling with network regularization,” in *Proceedings of the 17th International Conference on World Wide Web*, Beijing, China, 2008, pp. 101-110.



**Hyon Hee Kim** https://orcid.org/0000-0002-7507-8342

She received B.S., M.S., and Ph.D. degrees in Computer Science and Engineering from Ewha Womans University in 1996, 1998, and 2005, respectively. She joined the faculty of the Department of Statistics and Information Science at Dongduk Women's University, Seoul, Korea, in 2006. Her current research interests include machine learning, big data analysis, and ontologies.

**Hey Young Rhee** https://orcid.org/0000-0002-7701-1854

She received M.S. and Ph.D. degrees in Library and Information from Chung-Ang University in 2000 and 2009, respectively. She joined the faculty of the Department of Library and Information Science at Dongduk Women's University, Seoul, Korea, in 2014. Her current research interests include big data analysis.