kim_2019_an_ontology_based_labeling_of_influential_topics_using_topic_network_analysis

Year

2019

Author(s)

Hyon Hee Kim and Hey Young Rhee

Title

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

Venue

Journal of Information Processing Systems

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel approach

Underlying technique

Ontology-based labeling framework

Topic labeling parameters

Label generation

We proposed an ontology-based labeling framework. The UniDM ontology defines class

hierarchies and relationships among classes based on Wikipedia category.
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.

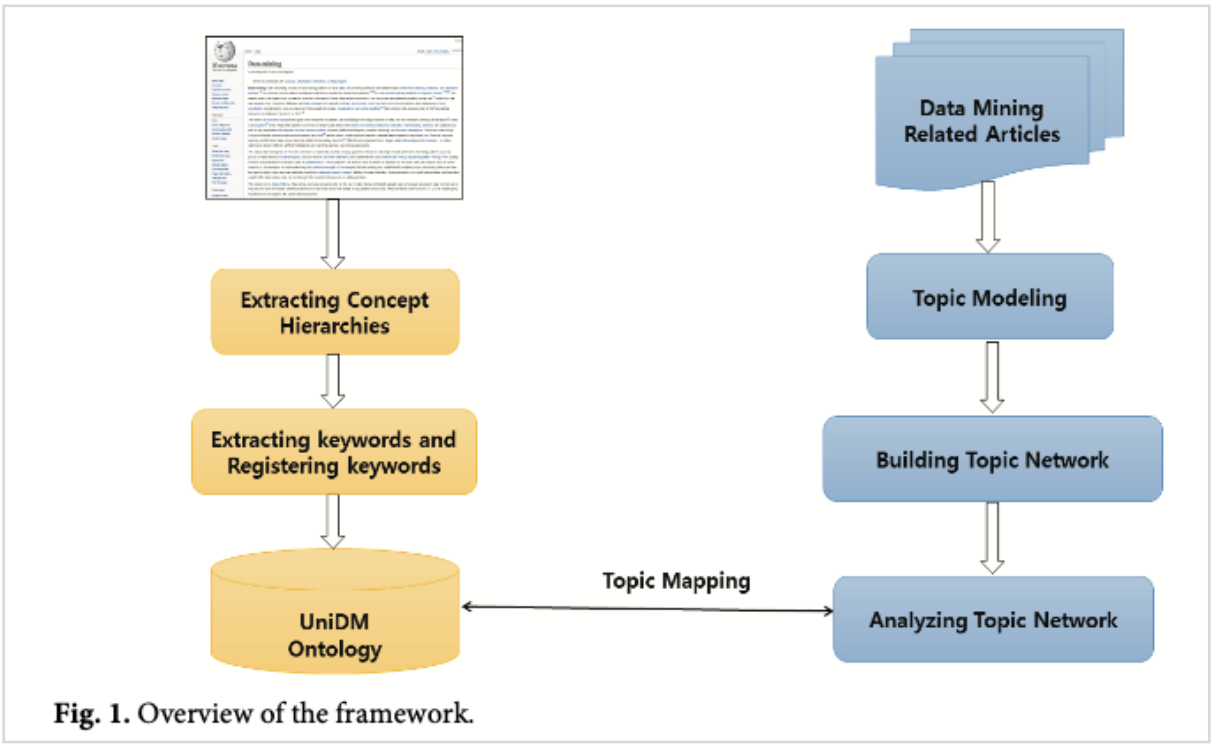

Fig. 1. Overview of the framework.

Side note:

- 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.
- As a result of the topic network analysis, influential topics are selected (Topics with the highest degree centrality, betweenness centrality, and closeness centrality), 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.

The 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

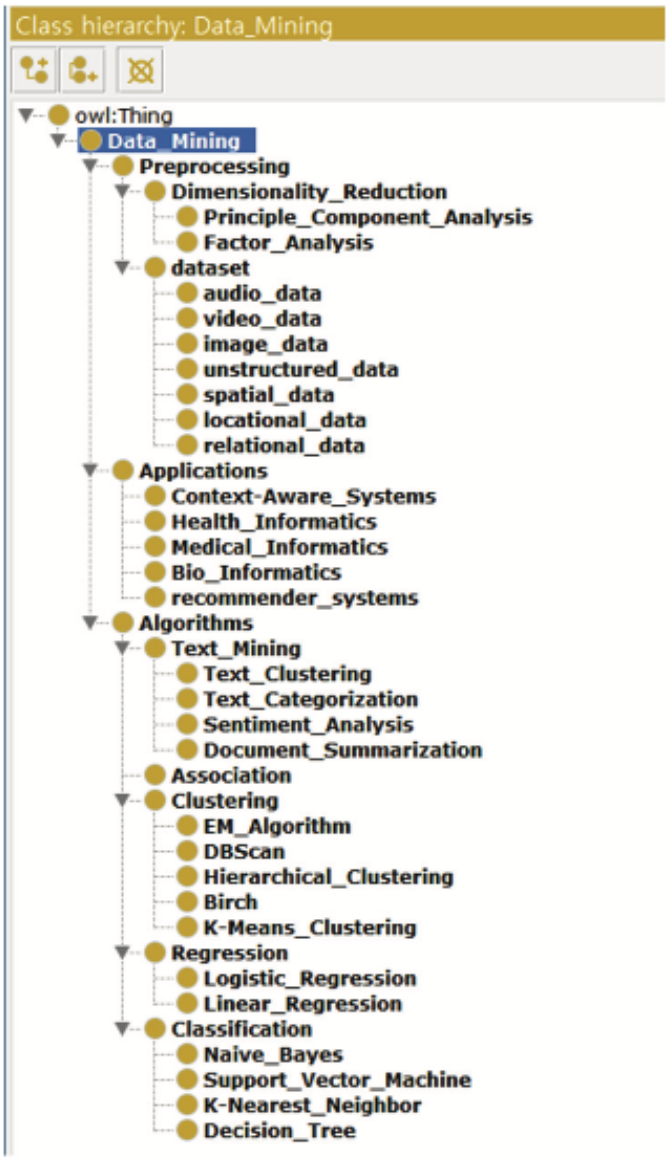

Fig. 2. Class Hierarchy in UniDM ontology.

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.

Results

Table 2 shows top 30 keywords belonging to the topic with the highest value of degree centrality, the topic with the highest value of betweenness centrality and 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

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

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.

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

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 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

Motivation

interpreting topics in topic models […] plays an important role in overcoming ambiguous and arbitrary interpretation of topics in topic modeling.

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.

Topic modeling

LDA

Topic modeling parameters

Nr of topics: {10, 20, 30, 40}

Nr. of topics

30

Label

Single or multi word label belonging to an instance of the classes in the UniDM ontology.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Topic labeling
Dataset: Data mining

Problem statement

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. Second, to interpret and to explain selected influential topics, the UniDM ontology is constructed from Wikipedia and serves as concept hierarchies of topic models.
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.

Corpus

Origin: Various computer science journals
Nr. of documents: 2,103
Details:

- Abstracts of research papers related to data mining published over the 20 years from 1995 to 2015 are collected and analyzed in this research.
- abstracts of “data mining” related articles from top 5 computer journals (*IEEE Transactions on Knowledge and Data Engineering*, *Data Mining and Knowledge Discovery*, *Information Sciences*, *Very Large Data Bases*, and *Expert Systems with Applications*)

Document

Abstracts of a research paper

Pre-processing

```
@article{kim_2019_an_ontology_based_labeling_of_influential_topics_using_topic_
network_analysis,
  title={An Ontology-Based Labeling of Influential Topics Using Topic Network
Analysis},
  author={Hyon Hee Kim and Hey Young Rhee},
  journal={J. Inf. Process. Syst.},
  year={2019},
  volume={15},
  pages={1096-1107}
}
```

#Thesis/Papers/BS