TOPIC MODELING AND ASSOCIATION RULE MINING TO DISCOVER GEOSPATIAL SEMANTIC INFORMATION FROM UNSTRUCTURED DATA SOURCES

E. Katsadaki[1]*, M. Kokla[1]

[1] School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Athens, Greece
(eirini.katsadaki@gmail.com, mkokla@survey.ntua.gr)

Commission IV, WG IV/2

KEY WORDS: Geospatial Knowledge, Semantic Information, Unstructured Data, Topic Modeling, Association Rules, Latent Dirichlet Allocation, FP-Growth

ABSTRACT:

As the amount of semi-structured and unstructured information sources expands at an exponential rate, there is a growing demand for semantic information elicitation of the immanent knowledge included in these sources. Semantic information elicitation processes such as semantic information extraction, linking, and annotation aim to make the knowledge explicit and unveil aspects latent in these sources to support knowledge discovery, semantic analysis, and visualization. The paper describes the implementation of Latent Dirichlet Allocation (LDA) topic modeling and association rule mining with FP-Growth for knowledge discovery. RapidMiner, an open-source data mining software is used for the objectives of this work.

1. INTRODUCTION

Exploring patterns within geospatial data is a concern for geographers, economists, and regional scientists now more than ever. Geospatial data contains information on the geographical location and characteristics of natural or artificial features and boundaries on the surface of the earth (Order, 1994). Geospatial data mining has been one of the most active areas of research in recent decades, owing to the tremendous expansion of the size of geospatial data. The purpose of geographic data mining is to uncover previously unknown non-trivial patterns. A spatial pattern that opposes randomness and causation can be a frequent arrangement, regularity, major direction, forecast, or composition. Geospatial data mining has a wide range of applications. The purpose of topic modeling, for example, is to organize texts based on term co-occurrences. Some tasks require the discovery of association rules that may be used to link events that are expected to occur together or occurrences that are ordered in time. Other tasks, such as developing regression models for a time series, focus on the temporal patterns of geospatial data.

Although geospatial data are commonly organized in structured data sources, research has recently also focused on the wealth of semi-structured and unstructured sources for extracting semantic information for geospatial concepts and entities. To facilitate knowledge discovery, semantic analysis, and visualization, semantic information elicitation techniques such as semantic information extraction, linking, and annotation strive to make knowledge explicit and reveal elements implicit in these sources (Kokla & Guilbert, 2020).

The paper describes the implementation of Latent Dirichlet Allocation (LDA) topic modeling and association rule mining with FP-Growth for knowledge discovery. RapidMiner, an open-source data mining software is used for the objectives of this work. The implementation of these techniques is demonstrated on a text collection derived from the book 'World Regional Geography' (World Regional Geography, 2016). The paper further discusses different parameters of each method that generate different results, the combination of topic modeling and association rules to identify term relationships, and how the results might be interpreted in a meaningful way to understand and explore emerging patterns.

The paper is organized as follows. Section 2 reviews previous work. Section 3 introduces the main methods used in this research. Firstly, Topic Modeling is introduced as an effective method for discovering topics in unstructured data sources. Then Association Rule Mining is presented in order to find meaningful and interesting associations between terms. Section 4 presents a synthesis of the results and discusses how the results might be interpreted in a meaningful way to understand and explore emerging patterns. Finally, Section 5 concludes our work and discusses future extensions.

2. RELATED WORK

Several approaches for extracting information from texts have been proposed for various subjects (Gangemi, 2013, Ristoski & Paulheim, 2016, Upadhyay & Fujii, 2016).

In the geospatial domain, among other semantic information extraction techniques, topic modeling has been successfully implemented to reveal latent abstract topics in text collections (Adams et al., 2015) or (Hu et al., 2017). Topic modeling such as Latent Dirichlet Allocation (LDA) is a text mining method for automatically detecting clusters of frequently co-occurring terms in a given text collection that represent abstract topics (Blei, 2012). However, although topic modeling techniques may describe higher-level semantic topics that characterize a text collection, they do not extract more intricate semantic

* Corresponding author

associations that may exist between terms that represent concepts and entities.
The acquisition of domain knowledge in the form of rules is a difficult undertaking due to the substantially more intricate logical links that must be modeled to acquire the rules (Kang & Lee, 2005). These techniques must automatically associate a sequence of such links to construct composite specifications that may be represented as rules, rather than finding correlations between pairs of ideas (Augier et al., 1995, Aharon et al., 2010, Schoenmackers et al., 2010).

Another promising text mining technique is Association rule mining. Association rule mining approaches are used by telecommunication networks, market and risk management, inventory control, bioinformatics, and other sectors (Dave et al., 2014) to identify intriguing correlations, patterns, relationships, and random structures between sets of items in large databases or other data repositories. Apriori and Frequent Pattern Growth (FP- Growth) (Han et al., 2004) are the most widely used algorithms for association mining.

The present work aims to delve deeper into revealing semantic patterns that represent relations between spatial concepts and entities in unstructured text collections. To accomplish this, topic modeling is combined with association rule mining to extract associations between spatial concepts and entities. So far, there has been a great number of results published by researchers regarding Topic Modeling (Gharbi et al., 2016), (Patil & Mante, 2018) separately or in combination (Dave et al., 2014) with Association rule mining mostly using structured data as input. The purpose of this study is to extract semantic information using unstructured data as input, for example, articles and books, to achieve the desired goals.

3. METHODOLOGY

This Section presents the methodology for extracting semantic patterns that represent relations between spatial concepts and entities from unstructured text collections. Figure 1 shows the process followed by the proposed approach. LDA is used to identify a set of abstract topics that are formed by terms that frequently co-occur and describe the subjects covered therein. Then, association rule mining is used to uncover more fine-grained connections between terms that represent spatial concepts and entities. These processes are performed using RapidMiner[1], an open source data science software platform for data preparation, machine learning, deep learning, text mining, and predictive analytics. The implementation of these techniques is demonstrated on a text collection derived from the book 'World Regional Geography' (*World Regional Geography*, 2016).

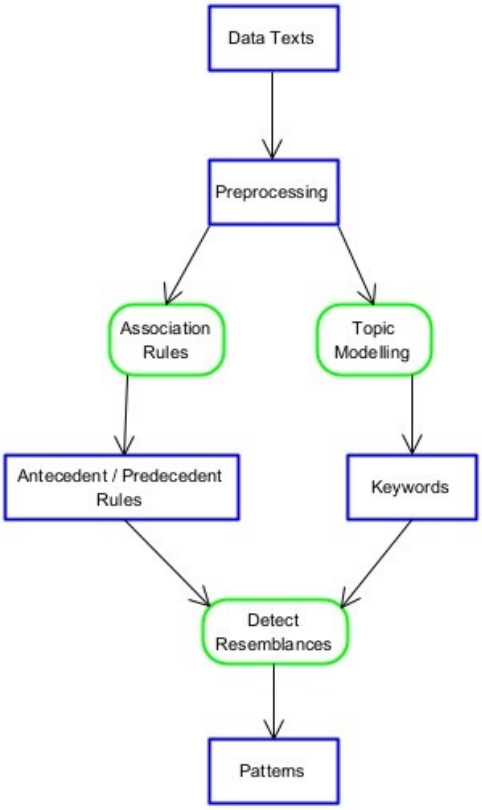

Figure 1: The proposed process that combines topic modeling and association rule mining to extract semantic information

3.1 Data Pre-processing

During the pre-processing step, the original unstructured data are converted into word vectors in order to enable the subsequent application of the various data mining methods.

The "Process Documents from Files" operator handles text processing, which comprises preparing text data for use with typical data mining techniques. This operator reads data from a set of text files and manipulates it using text processing algorithms. This is a nested operator, which means it contains a sub-process made up of six serially linked operators (Figure 2):

- Tokenize Non-letters (Tokenize)
- Tokenize Linguistic (Tokenize)
- Filter Stopwords (English)
- Filter Tokens (by Length)
- Stem (Porter)
- Transform Cases

The sub-process essentially converts text data into a format that can be easily studied using standard data mining techniques like association rule mining and topic modeling.

The Tokenize Non-letters (Tokenize) and Tokenize Linguistic (Tokenize) operators are both formed in this sub-process by selecting the Tokenize operator, but with distinct parameter options. The former operator tokenizes based on non-letters, whereas the later operator tokenizes based on English language linguistic phrases.

[1] https://rapidminer.com/

The Filter Stopwords (English) operator removes stop words from the text data set in the English language.

The Filter Tokens (by Length) operator filters out all terms on the basis of predefined min and max characters.

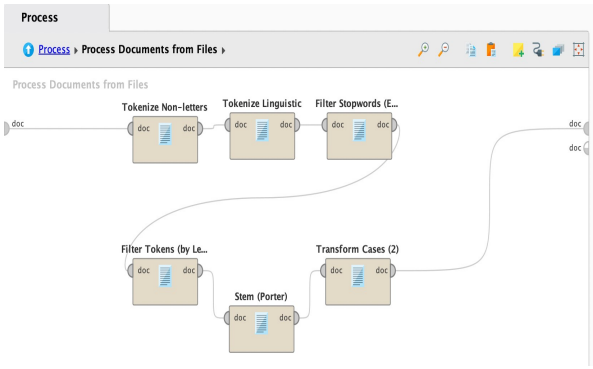

Figure 2. Pre-processing Documents from Files

Next, Term Frequency, Term Occurrences, and Term Frequency-inverted document frequency (TF-IDF) are three methods used for transforming terms into vectors. The term vector provides quantitative values for each term in a document, which is extremely useful in feature selection. TF-IDF is the most helpful and widely used (Jing et al., 2002). This provides the important term a higher weighting and the unimportant term a lower weighting. Its vector value ranges from 0 to 1. 0 indicates that the phrase has no meaning in the context of the documents we're looking for, whereas 1 indicates that the terms are significant. To compute the TF-IDF values, the document is converted into an inverted text file, in which all documents' words are removed and a weight is assigned to each word, i.e. the term occurrence value and document occurrences are calculated using the TF-IDF value.

All of the preprocessing steps listed above were used in both processes, and the TF-IDF approach was used to create the terms' vector. The Text to Nominal operator converts text into numerical (categorical) data. The data is subsequently transformed into binomial form using the Numerical to Binomial operation.

3.2 Topic Modeling

The two major goals of Topic Modeling are to find hidden themes in text data by clustering related words and then classifying the document into the found themes. Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA) are the two most common topic models. PLSI is a technique that uses matrix decomposition to discover latent topics. LDA, on the other hand, is a probabilistic generative corpus model that employs Dirichlet over the latent topic. The core idea is that documents are represented as random mixtures of latent subjects, each of which is described by a word distribution (Ankarali et al., 2020). The PLSI is easier to train than the LDA, but it has lower accuracy, which is why we chose the LDA for our process.

3.3 Process of Topic Modeling

The Operator Toolbox module in RapidMiner is used for topic modeling, and the LDA approach is used to uncover latent themes in processed data. The data are set to iterate 10 times because topic modeling is unsupervised learning. The technique of extracting topics from a group of documents using LDA is depicted in Figure 3. Before being clustered into a similar group, all data is pre-processed. The procedure of topic annotation is then carried out in order to organize the topics into a coherent theme. The number of topics was set to 4 after different trials.

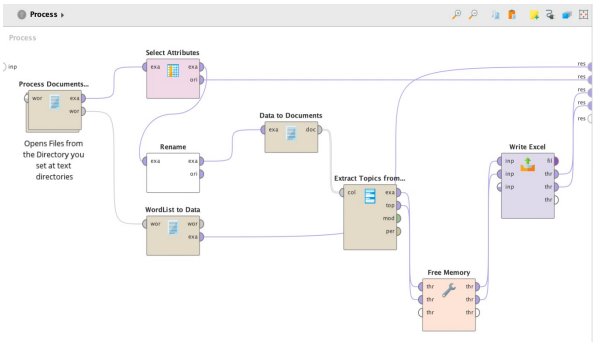

Figure 3. Topic Modeling Process on RapidMiner

3.4 Association Rule Mining

The technique of association rule mining is used to discover relations between hidden patterns in massive databases. The study's main goal is to find interesting associations, common patterns, or informal structured collections.

The FP-growth algorithm navigates from the bottom up to build frequent data sets from the FP Tree (Abdirad & Mathur, 2021). By constructing a condensed type of the data source in terms of an FP Tree, this approach reduces the total number of user data sets. The successful discovery of frequent data sets is enabled by this frequent information. This is a two-step process that's more efficient than previous association mining techniques (Jian-wen et al., 2008).

Step 1. Creates the FP Tree, a compact user navigation system. It's made up of two passes through the data collection.
Step 2. Extracts frequent set items from FP-tree Traversal through FP-Tree.

A basic operator workflow is depicted in Figure 4. The model is applied to the complete dataset.

The vector generating method employed is a critical parameter. As previously stated, the chosen vector generating method is TF-IDF. However, this may result in an excessive number of words, possibly tens of thousands; therefore a prune method is used to prune the resulting word set. As can be seen from the setting of 30.0 for the prune below percent option in ARM Process, words that appear in less than 30.0 percent of the documents are pruned.

The minimum support argument that was chosen in our experiments is 0.5, which means that the operator generates a list of the frequently occurring collections of words (item sets) that exist in at least 50% of the documents. Furthermore, because the max items option is set to 2, the resulting list is

limited to pairs of words (2-itemsets), and it does not contain frequent word sets (item sets) with three or more words. The FP-Growth operator passes the list of frequent word sets to the Create Association Rules operator, which computes the rules that satisfy the stated constraints on selected association mining criteria. The association rules are computed in our experiments using the confidence criterion, as well as gain theta and laplace k.

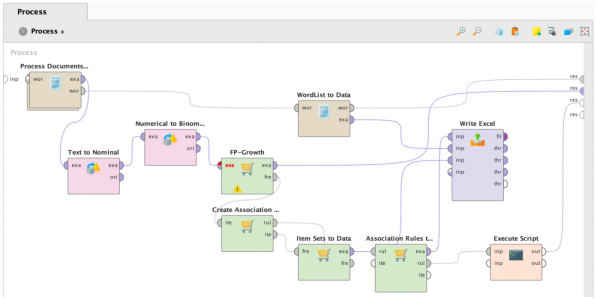

Figure 4. Association Rules Mining Process on RapidMiner

4. SYNTHESIS OF RESULTS

As mentioned above, the optimal topic allocation after the application of topic modeling was 4. Each topic brings out the 20 most important words regarding the meaning of each topic. The word clouds which illustrate the top 20 keywords related to the respective topics are presented in Figure 6.

Figure 5 shows an example for the association rules generated for the word "Asia".

Table 1 presents the most indicative association rules detected after running ARM process. The rules presented are based on the support parameter and were selected based on geographical, social, and economic criteria. Our main goal in this paper is to discuss a possible combination of the results of each process and to observe the existence of a pattern between the topics and the most important association rules.

Indicatively, we will check the most important words of topic 3 and investigate the existence of association rules found by the tool.

For example, a highly weighted term in topic 3 is the term "Asia" (Figure 6). Topic 3 also contains other meaningful terms related to the term Asia such as the terms China, Pakistan, Russia, India, Japan which represent a geospatial connection with Asia and the terms communist, soviet, economy and industrial which represent social, economic and political aspects.

The association rules for the term Asia which can be found in all the chapters with support 92.9% are:

1. Asia -> wall (support 60%)
2. Asia -> industrial (support 50%)
3. Asia -> tropic (support 50%)

Then we can observe that the rules for the words wall and tropic are connected with the rules for word Asia and consist of a pattern:

1. Wall -> border (support 50%)
2. Wall -> China (support 50%)
3. Tropic -> India (support 50%)
4. Tropic -> island (support 50%)
5. Tropic -> land (support 50%)
6. Tropic -> population (support 50%)
7. Tropic -> region (support 50%)
8. Tropic -> tourism (support 50%)
9. Tropic -> travel (support 50%)
10. Tropic -> landscape (support 50%)
11. Tropic -> rain (support 50%)
12. Tropic -> terrain (support 50%)
13. Tropic -> species (support 50%)

The association rules for word Asia and wall results to a logical pattern which represents the term Asia which is one of the most frequently encountered word based on the support parameter. It was observed that the next word with the biggest support is the word wall. For this reason, all the association rules between these two words were studied as they constitute a two-way connection. Then, we observe the terms industrial and tropic which are highly associated with the word Asia as well. The words wall and tropic are also linked with the words border, China, India, island, land, population, region, tourism, travel, landscape, rain, terrain, and species as it is presented on Table 1. These words consist of an expected and meaningful set of words which are clearly associated with the word Asia based on geographical, social and spatial aspects. The word industrial was not further investigated because no association rule emerged with support greater than 50% as set in the RapidMiner parameters. Figure 7 presents the associations' diagram of our example for the word "Asia".

In conclusion, it was observed that the most important words according to topic modeling, also formed strong association rules with support over 50% that were interconnected at multiple levels, as was explained in the example with the word Asia.

Size	Support	Item 1	Item 2
1	0.929	Asia	
1	0.700	wall	
2	0.600	Asia	wall
2	0.500	Asia	industrial
2	0.500	Asia	tropic
2	0.500	border	wall
2	0.500	India	tropic
2	0.500	island	tropic
2	0.500	land	tropic
2	0.500	population	tropic
2	0.500	region	tropic
2	0.500	tourism	tropic
2	0.500	travel	tropic
2	0.500	landscape	tropic
2	0.500	rain	tropic
2	0.500	terrain	tropic
2	0.500	species	tropic
2	0.500	China	wall

Table 1. Most Indicative Association Rules

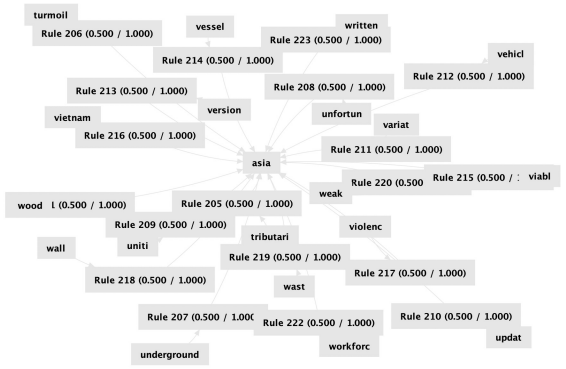

Figure 5. Visualization of Association Rules in RapidMiner

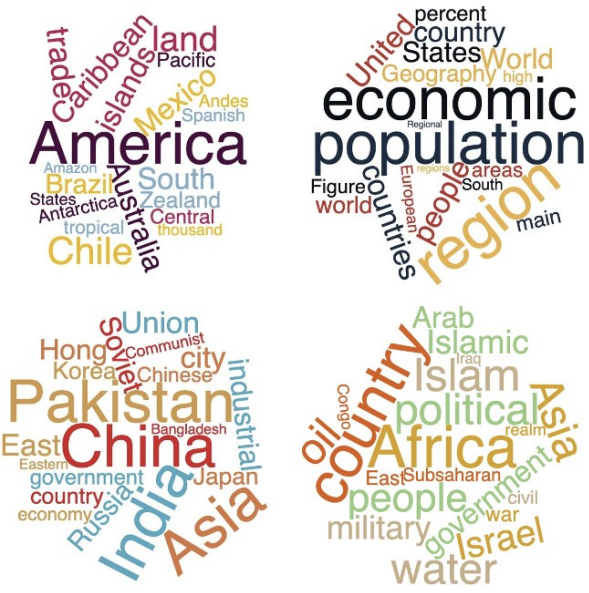

Figure 6. Word clouds with the top 20 keywords related to the 4 respective topics.

5. CONCLUSIONS

The paper describes the application of topic modeling and association rule mining to explore interesting patterns and associations between spatial concepts and entities in unstructured data sources.

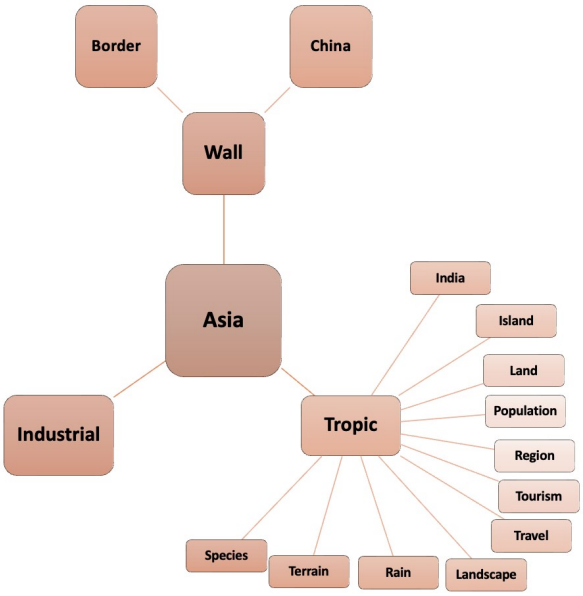

Figure 7. Associations Diagram for the word "Asia".

A multifaceted framework is developed for extracting knowledge from unstructured data sources and converting it into logical forms utilizing text mining and natural language processing (NLP) methods. The combined results of Topic Modeling and Association Rule Mining revealed significant patterns that could support processes such as ontology enrichment and semantic annotation.

Although machine learning methods are becoming mainstream for interpreting geospatial information, there is a need to release data and source code publicly, to create and maintain credible benchmarking activities, and to quantitatively test algorithms on open large scale datasets in order to further accelerate research.

When constructing and implementing an evolutionary ARM or TM algorithm, a number of factors must be taken into account. Large datasets, attribute values, and parameter settings are all issues that affect both development and application. One of the most significant obstacles to implementing metaheuristic algorithms in ARM is that determining appropriate parameters, such as minimal support and confidence criteria requires substantial skill and experience. The quality of patterns recovered by the ARM and TM processes is closely related to these factors. The challenges define the work that needs to be done in the future to better text mining of items.

Data heterogeneity is one major challenge for text mining, further complicated by the use of several languages and the range of notations used for the same word. Another obstacle is that sparse data leads to data overfitting into clusters and classifications, resulting in incorrect analysis. Domain expertise is essential for providing proper analysis for textual data, and having a qualified domain expert may not always be attainable.

REFERENCES

Abdirad, H., & Mathur, P. (2021). Artificial intelligence for BIM content management and delivery: Case study of association rule mining for construction detailing. *Advanced*

Engineering Informatics, *50*, 101414. https://doi.org/10.1016/j.aei.2021.101414

Adams, B., McKenzie, G., & Gahegan, M. (2015). Frankenplace: Interactive Thematic Mapping for Ad Hoc Exploratory Search. *Proceedings of the 24th International Conference on World Wide Web*, 12–22. https://doi.org/10.1145/2736277.2741137

Aharon, R., Szpektor, I., & Dagan, I. (2010). Generating Entailment Rules from FrameNet. *Proceedings of the ACL 2010 Conference Short Papers*, 241–246. https://aclanthology.org/P10-2045

Ankarali, E., Külcü, Ö., & Kl, J. (2020). Topic Modeling of Twitter Data via RapidMiner. *Bilgi Yönetimi*, *3*, 1–10. https://doi.org/10.33721/by.641878
Augier, S., Venturini, G., & Kodratoff, Y. (1995). *Learning First Order Logic Rules with a Genetic Algorithm*. 21–26.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Dave, N., Davis, D., Potts, K., & Asuncion, H. U. (2014). Uncovering file relationships using association mining and topic modeling. *The Sixth International Conference on Information, Process, and Knowledge Management*, 105–111.

Gangemi, A. (2013). A Comparison of Knowledge Extraction Tools for the Semantic Web. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, & S. Rudolph (Eds.), *The Semantic Web: Semantics and Big Data* (Vol. 7882, pp. 351–366). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38288-8_24

Gharbi, A., De Runz, C., Faiz, S., & Akdag, H. (2016). Towards Association Rules as a Predictive Tool for Geospatial Areas Evolution: *Proceedings of the 2nd International Conference on Geographical Information Systems Theory, Applications and Management*, 201–206. https://doi.org/10.5220/0005914202010206

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery*, *8*(1), 53–87. https://doi.org/10.1023/B:DAMI.0000005258.31418.83

Hu, Y., Ye, X., & Shaw, S.-L. (2017). Extracting and analyzing semantic relatedness between cities using news articles. *International Journal of Geographical Information Science*, *31*(12), 2427–2451. https://doi.org/10.1080/13658816.2017.1367797

Jian-wen, X., Yan-xiang, H., Futatsugi, K., & Wei-qiang, K. (2008). Constructing projection frequent pattern tree for efficient mining. *Wuhan University Journal of Natural Sciences*. https://doi.org/10.1007/BF02907210

Jing, L., Huang, H., & Shi, H. (2002). Improved feature selection approach TFIDF in text mining. *Proceedings. International Conference on Machine Learning and Cybernetics*. https://doi.org/10.1109/ICMLC.2002.1174522

Kang, J., & Lee, J. K. (2005). Rule identification from Web pages by the XRML approach. *Decision Support Systems*, *41*(1), 205–227. https://doi.org/10.1016/j.dss.2005.01.004

Kokla, M., & Guilbert, E. (2020). A Review of Geospatial Semantic Information Modeling and Elicitation Approaches. *ISPRS International Journal of Geo-Information*, *9*(3), 146. https://doi.org/10.3390/ijgi9030146

Order. (1994). Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure. *Federal Register*, *59*(71), 4.

Patil, P. P., & Mante, P. R. V. (2018). Discovering Interesting Locations in a Geospatial Region using Association Rule Mining. *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. https://doi.org/10.1109/RTEICT42901.2018.9012351

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *J. Web Semant.* https://doi.org/10.1016/j.websem.2016.01.001

Schoenmackers, S., Davis, J., Etzioni, O., & Weld, D. (2010). Learning First-Order Horn Clauses from Web Text. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1088–1098.

Upadhyay, R., & Fujii, A. (2016). *Semantic Knowledge Extraction from Research Documents*. 439–445. https://doi.org/10.15439/2016F221

World Regional Geography: People, Places and Globalization (01 ed.). (2016). University of Minnesota Libraries Publishing. https://doi.org/10.24926/8668.2701