A vocabulary recommendation method for spatiotemporal data discovery based on Bayesian network and ontologies

Kejin Cui¹, Yongyao Jiang¹,², Yun Li³ and Dieter Pfoser¹

¹Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA; ²Environmental Systems Research Institute, Redlands, CA, USA

ABSTRACT

In the research field of spatiotemporal data discovery, how to utilize the semantic characteristics of spatiotemporal datasets is an important topic. This paper presented a content-based recommendation method, and applied Bayesian networks and ontologies into the vocabulary recommendation process for spatiotemporal data discovery. The source data of this research was from the MUDROD (Mining and Utilizing Dataset Relevancy from Oceanographic Datasets) search platform. From the historical search log, major keywords were extracted and organized according to ontologies in a hierarchical structure. Using the search history, the posterior probability between each subclass and their super class in the ontologies was calculated, indicating a recommendation likelihood. We created a Bayesian network model for inference based on ontologies. This model can address the following two objectives: (1) Given one class in the ontology, the model can judge which class has the biggest likelihood to be selected for recommendation. 2) Based on the search history of a user, the Bayesian network model can judge which class has the biggest probability to be recommended. Comparison experimentation with existing system and evaluation experimentation with expert knowledge show that this method is specifically helpful for spatiotemporal data discovery.

1. Introduction

Spatiotemporal big data can be characterized by 5Vs: volume, velocity, variety, veracity and value (Yang, Huang, Li, Liu, & Hu, 2017). With these characteristics being interconnected, the challenge to transform Big Data’s first 4Vs into the 5th (value) requires novel scientific research, engineering, and business intelligence (Lee & Kang, 2015). To better manage and utilize geospatial data, the concept of spatial data infrastructure (SDI) was introduced and applied in GlScience discipline (Yang, Raskin, Goodchild, & Gahegan, 2010), and multiple geoportals were built for big geospatial data management by various organizations and countries (Maguire & Longley, 2005). Examples of such geoportal websites are NASA’s data portal (Edberg et al., 2016), GEOSS Portal by Group on Earth Observations (Giuliani et al., 2013), INSPIRE by European Union (Bernard, Kanellopoulos, Annoni, & Smits, 2005), and data.gov by the US government (Hendler, Holm, Musialek, & Thomas, 2012). The common
challenge in all these efforts is to design a model for efficient data discovery, which would make a significant contribution to such data infrastructure efforts.

Part of the challenge here is that spatiotemporal data comes from different sources, involving different data formats, and different temporal and spatial granularities. Heterogeneous spatiotemporal data is inevitable (Bogdanović, Stanić, & Stoimenov, 2015). This heterogeneity makes it hard for the user base of such data portals to search for and obtain the data they are looking for. The users of SDI need a mechanism to easily find (discover) what they are searching for – ideally using their own vocabulary and language (Tait, 2005). Therefore, research in smart geospatial information discovery is designing methods based on data semantics (relates to language) and artificial intelligence to facilitate the retrieval and access of geospatial information based on the users’ search demand. Improved search and discovery mechanisms will also increase the value of spatiotemporal big data research.

This paper presents a new method for spatiotemporal data discovery based on Bayesian networks and ontologies, which is an important semantics-based method in artificial intelligence. Compared to existing approaches, this semantics-based method will show significant application value in effective spatiotemporal data discovery. Section 2 presents a review of related approaches. Section 3 discusses the general research method of this paper. Section 4 outlines the process model creation, including the creation of ontologies, keyword extraction, and the creation of the Bayesian network model. Section 5 compares the performance of our method to existing data discovery portals, and evaluate our method based on expert knowledge. Section 6 concludes and presents directions for future work.

2. Literature review

A considerable effort has been dedicated to the design of recommendation systems (Aggarwal, 2016). Existing recommendation approaches include collaborative filtering, content-based filtering and hybrid recommendation approaches (Asanov, 2011). In existing spatiotemporal data discovery systems, collaborative filtering has been applied by considering the different users’ search, click and download behaviors in relation to a specific dataset (Jiang et al., 2017). This method is mainly focused on user’s behaviors, their interactions and similarity, so it is generally user-centered. However, the users of earth observation data search engines are typically researchers, which is a small and specialized user group when compared to the case of traditional and considerably larger recommendation systems. Hence, geospatial data portals are typically confronted with “cold starts” with many datasets having few or even zero clicks or download history. Here, a user-centered recommendation method would be not very helpful. On the other hand, when adding new datasets to the system, this update may not be quickly reflected in the recommendation results based on collaborative filtering methods. To resolve this “cold start” problem, content-based methods will play a more important role, which is the main focus of this paper.

In previous work, most SDI and data portals rely on keyword-based searches, and at the core of these techniques is Apache Lucene, an open source information retrieval tool (McCandless, Hatcher, & Gospodnetic, 2010). Lucene provides powerful full-text search capabilities but fails to address the “intent gap” (i.e., the gap between its representation of users’ queries vs. their true intent) (Jiang et al., 2017). By combining semantically related terms, query modification has been proposed as a solution to bridge this gap (Hua et al.,
2013; Mangold, 2007). For example, “precipitation amount” and “rainfall amount” are related terms, so it is appropriate to combine these two semantically related terms to modify search result. On the other hand, this paper will propose another alternative method to resolve the semantic gap between users’ queries and their true intent. For instance, when searching “ocean” and “wind”, data portal users may not find their satisfying datasets. If we could propose a model to recommend “NetCDF” or “surface” to them, probably they could add these two keywords in their queries and find the exact dataset they need.

In artificial intelligence, the semantics of data has emerged as a common, affordable concept that facilitates interoperability, integration and monitoring of knowledge-base systems (Kourtesis, Alvarez-Rodríguez, & Paraskakis, 2014). The process of data discovery in an SDI is a multi-step process, starting with the users’ task to formulate their information need as a query, processing of the query, finding and returning matching metadata in the repository, and finally evaluating the results, so semantics can help in each of these steps (Janowicz et al., 2010). Hence, utilizing the semantic characteristics of spatiotemporal data will be very helpful to improve the efficiency of data discovery. For example, Liu (2011) improved the recall and precision of geospatial metadata discovery based on Semantic Web for Earth and Environmental Terminology (SWEET) ontology, concept and relationships. (Jiang et al., 2017) presented a comprehensive methodology based on semantic similarity analysis and ontologies to improve data relevancy of smart search for oceanographic data discovery.

Ontology, is originally a research discipline in philosophy, which explores existence (Smith, 2003). The ideas of ontologies have been widely applied in information science, especially semantic web, as a tool for data modeling (Maedche & Staab, 2001). Ontologies have been applied in recommendation system to capture the knowledge of user preferences (Middleton, De Roure, & Shadbolt, 2001). One of the classic applications of ontologies in information science is OWL, which is a web language based on ontologies and can be applied in information query and process. On the other hand, in operational languages of ontologies, a probabilistic reasoning language, called PR-OWL (Carvalho, Laskey, & Costa, 2010; Da Costa, Laskey, & Laskey, 2008), was presented to perform probabilistic inferences in ontology-based models. In spatiotemporal data discovery, ontologies play an important role in metadata modeling of semantics under global and ubiquitous computing environment (Pfoser, Pitoura & Tryfona, 2002). Several semantics-based services have been built to apply ontologies in spatial data infrastructure (Lacasta, Nogueras-Iso, Béjar, Muro-Medrano, & Zarazaga-Soria, 2007; Li et al., 2011). Rule-based ontologies with reasoning and description logic have been applied in spatial data infrastructure (Lutz & Kolas, 2007). In existing research on spatiotemporal data discovery, ontologies were mainly applied in conceptual modeling of knowledge engineering, but the perspective of probability inference and reasoning was not sufficient (Cui & Bittner, 2016). Our paper focusses on this gap to perform probabilistic inference based on ontologies to improve spatiotemporal data discovery.

3. Research method

Vocabulary recommendation is an important function for a spatiotemporal data infrastructure. By discovering semantic relationships among keywords in a data portal, more useful keywords could be recommended to users to help them find the dataset they really intend to find. The present effort towards spatiotemporal data discovery is
aimed at utilizing the semantic characteristics of datasets. We combine Bayesian networks and ontologies towards improving recommendations in spatiotemporal data discovery.

Bayesian networks (Jensen, 1996) – also known as “belief networks” or “causal networks” – are graphical models for representing multivariate probability distributions. Each variable $X_i$ is represented as a vertex in a Directed Acyclic Graph (DAG); the probability distribution $P(X_1, X_2, \ldots, X_n)$ is represented in factorized form as follows:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \prod X_i)$$

where $\prod X_i$ is the set of vertices that are $X_i$’s parents in the graph.

In a Bayesian network model, each edge of the network indicates posterior probability from one entity to another.

We choose and combine these two methods for spatiotemporal data discovery for the following reasons. First, the method we choose is a content-based recommendation method. As mentioned, a search engine for scientific data has to address the “cold start” problem, i.e., datasets are specialized and do not have a mass appeal resulting in a large number of search, and sometimes we have to add some new datasets so they may not have many search or download records. However, existing method based on collaborative filtering cannot deal with this problem well, because it only considers specific users’ search history and their interactions. Here, content-based methods will be more reliable when dealing with such “cold start” problems as they do not rely on users’ specific behavior. Second, the method we propose is a semantics-based recommendation method. By using ontologies, keywords are organized based on their semantic relationship, and this relationship can be organized as a Bayesian network model for probabilistic inference. Third, the proposed model integrates conceptual modeling of ontologies with quantitative reasoning of a Bayesian network. The concept of ontologies was originally presented for conceptual modeling. On the other hand, the hierarchical characteristics of ontologies could be thought of as a DAG (Directed Acyclic Graph) model. This hierarchical network model characteristic can be integrated with a Bayesian network model to allow for probabilistic inference. This combined probabilistic reasoning model will provide an efficient means for content-based spatiotemporal data discovery.

The specific model implementation is as follows. Initially, based on the (limited) search history of users, the total search count of each keyword, i.e., the frequency with which each keyword has appeared in recommendations, is calculated. This creates the limited basis for the recommendation. Next, ontologies are created that organize these keywords in a hierarchical structure. The frequency of each word and its corresponding place in this structure form a Bayesian Network. Finally, the posterior probability between each super class and their subclasses is calculated, indicating the likelihood of a recommendation based on the search activities of users. The general workflow of research method is shown in Figure 1.

In general, this model will support the following two use cases:

1. Given one class in the model, it can automatically judge which class in the ontology has the biggest likelihood to be selected for recommendation.
(2) Based on the search history of a user, the system will create a Bayesian Network model to capture this search history. For each user, the system can provide the class in the ontology that has the biggest likelihood to be recommended.

4. Model creation

The data utilized in this research stems from the MUDROD (Mining and Utilizing Dataset Relevancy from Oceanographic Datasets to Improve Data Discovery) search platform. From the search log of this search platform, we can extract important keywords and their corresponding frequencies, by simply summing up their occurrences in searches. From the search log data, we can for example identify 20 most frequent keywords based on the users’ overall search history, as shown in Table 1.

Calculating the frequency for all keywords, we only consider those with frequency of >1000 for our model. This results 69 keywords for this specific dataset. Based on domain knowledge of spatiotemporal data, these keywords are now organized by a domain expert in an ontology as follows. Each vocabulary is judged and categorized under a subclass of the ontologies.

As shown in Figure 2, there are nine major classes in this ontology: OCEAN_ENTITY, VIARABLE, RESOLUTION, DATA_MODEL, GROUP, SENSOR, ACTION, LOCATION and ADJECTIVE.

OCEAN_ENTITY: Geographic phenomena related to the concept of ocean.
VIARABLE: Variables to describe ocean entities.
RESOLUTION: Including resolution, the spatial and temporal interval for the datasets.
DATA_MODEL: Data models to describe ocean related data.
GROUP: Related groups to support ocean science study and research.
SENSOR: Including instruments and satellites. Sensors to collected oceanographic data.
ACTION: All operations to access ocean data.
LOCATION: The geographic location.
ADJECTIVE: Some adjectives to describe the dataset for the search process.

Next, all of the 69 keywords are categorized in this ontological structure and a Bayesian network is created. The posterior probability is calculated to describe the relationship between each subclass and its super class. A partition method is used to build out the Bayesian network. According to the characteristics of Bayesian network, the sum of the probability in each box component should be 100%, so an additional class “OTHER” appears

![Figure 2. The general structure of the ontologies in this project.](image-url)
in each box component of the Bayesian network model to satisfy this characteristics. The overall structure of this Bayesian Network model is shown in Figure 3.

Based on this Bayesian Network model, some interesting analysis can be done in support of a recommendation system. Some example requests are as follows.

Based on the general structure of this Bayesian Network model, all of the prior and posterior probability of each keyword (concept) and its super class can be obtained to summarize the search history. Three major upper classes are mostly searched: OCEAN_ENTITY (29.5%), RESOLUTION (27.8%) and SENSOR (15.0%). The most frequent data model is NETCDF, and the two most frequent searched satellites are JASON-1 and GRACE.

Given a specific keyword, for instance, WINDS, the five most related keywords with the biggest probability according to this Bayesian Network are: SEAWINDS, TEMPERATURE, DAY, GLOBAL and SURFACE.

Suppose a user never search any keywords in the OCEAN_ENTITY and RESOLUTION category, then the three most frequent categories are SENSOR, DATA_MODEL and VARIABLE. The first five keywords to be recommended should be: TEMPERATURE, NETCDF, AVHRR-3, GHRSSST, and RAW.

Suppose a user would like to search ATMOSPHERIC and VECTOR, and then the five keywords to be recommended should be: OCEAN, TEMPERATURE, DAY, GLOBAL, and HISTORICAL. Users could add any of these keywords to their search queries, and finally find their target datasets.

Figure 3. Bayesian network for data discovery.
5. Model evaluation

Having developed a Bayesian network model to provide recommendations for the case of spatiotemporal data discovery, we will in this section provide experimentation to show that the results are also satisfying and relevant. We propose two experiments to evaluate the performance of our approach. The first experiment is a comparison experiment between the proposed model and existing recommendation systems using based on precision and recall. The second experiment is an evaluation experiment that utilizes domain experts to assess result quality.

5.1. Existing recommendation system

In this first experiment we will compare the recommendations of our method to an existing recommendation system. Given that we are using the data of an existing system, it is only natural to compare the performance of our method to the recommendations of MUDROD portal. The system is available at https://mudrod.jpl.nasa.gov/#/. The recommendation method exploits the users’ search and click behavior based on a collaborate-filtering method (Jiang et al., 2017). The top 20 search results for each of the keywords in question will be extracted and their top 10 respective keywords will be summarized. We compare the existing method to our approach using precision and recall. Since the number of candidate keywords is 10 for both precision and recall calculation, which is the denominator for both of them, so the value of precision and recall is the same.

In this experiment, 5 keywords are chosen for comparison: WIND, NETCDF, OCEAN, GRACE and RADAR. The result summary is shown in Table 2.

From the table above, the top 10 keywords generated from search result and model inference are compared. From the result of precision and recall, it could be concluded that the proposed Bayesian model could identify the most frequent keywords when compared to the original data discovery system with acceptable errors. The results also show the proposed model could be an alternative content-based recommendation method.

| Search Keyword | Result (Search Result/Model Inference) | Precision | Recall |
|----------------|----------------------------------------|-----------|--------|
| WIND           | Native Result: OCEAN, SURFACE, SPATIAL RESOLUTION, ICE, QUICKSCAT, DAY, NETCDF, SEAWINDS, DATASET, GLOBAL Bayesian Model Inference: AVHRR-3, TEMPERATURE, DAY, GLOBAL, SURFACE, OCEAN, NETCDF, SPATIAL RESOLUTION, SEAWINDS, GHRSST | 70%        | 70%    |
| NETCDF         | Native Result: OCEAN, GRACE, SURFACE, GLOBAL, DAY, GHRSST, TEMPERATURE, JASON-1, ASCII, PRESSURE Bayesian Model Inference: OCEAN, SURFACE, WIND, GHRSST, TEMPERATURE, AVHRR-3, GLOBAL, JASON-1, HISTORICAL, DAY | 70%        | 70%    |
| OCEAN          | Native Result: GLOBAL, WIND, DAY, SURFACE, AVHRR-3, GHRSST, DATASET, MONTHLY, NETCDF, TEMPERATURE Bayesian Model Inference: GHRSST, AVHRR-3, TEMPERATURE, NETCDF, DATASET, HISTORICAL, DAY, GLOBAL, SURFACE, WIND | 90%        | 90%    |
| GRACE          | Native Result: WATER, OCEAN, GLOBAL, GRAVITY, MONTHLY, DATASET, TEMPERATURE, SPATIAL RESOLUTION, SURFACE, NETCDF Bayesian Model Inference: OCEAN, SURFACE, WIND, GHRSST, AVHRR-3, TEMPERATURE, GLOBAL, NETCDF, MONTHLY, SPATIAL RESOLUTION | 70%        | 70%    |
| RADAR          | Native Result: NETCDF, OCEAN, JASON-1, GLOBAL, SPATIAL RESOLUTION, SURFACE, DAY, SEAWINDS, RAPIDSCAT, SOUTHERN ASIA Bayesian Model Inference: SEAINDS, WINDS, OCEAN, SURFACE, DAY, NETCDF, GLOBAL, TEMPERATURE, SPATIAL RESOLUTION, JASON-1 | 80%        | 80%    |
5.2. Expert knowledge and recommendation quality

In this experiment, ocean scientists from NASA’s Jet Propulsion Laboratory evaluated the similarity between some sample metadata files. In total there are three sample query datasets, each of them includes 7–8 candidate search results with their corresponding evaluation for recommendation. This evaluation result was given in qualitative terms ranging from “Excellent”, “Good”, “OK” to “Bad”. Based on the expert knowledge above, we can design the comparison experiment as follows. Firstly, for each query metadata, find 5 keywords as “seed keywords” and then use our Bayesian model to infer 10 keywords, which are the most similar to these 5 seeds. Next, from the pool of recommendation result metadata and for each evaluation result group (“Excellent”, “Good”, “OK” and “Bad”), pick two metadata files and choose from each set of two files 15 keywords. Finally, calculate the precision between the sets of keywords, i.e., query and the recommendation results.

Tables 3–5 describes the three sample queries, with their 5 seed keywords and 10 inferred keywords from our proposed Bayesian network model. Next, for each evaluation, two sample recommendation metadata datasets are shown with their 15 keywords and the precision is calculated.

The experiments show that the proposed Bayesian network model could provide in many cases good recommendation results. We observe that when an evaluation by an expert drops from “Excellent” to “Bad”, the precision of the result also decreases. As such, our Bayesian model based recommendation mimics expert knowledge and provides accurate content-based recommendations for spatiotemporal data discovery.

6. Conclusion

This paper developed a novel recommendation system based on ontologies and the Bayesian Network technique for the case of spatiotemporal data discovery. In general, this model effectively organizes all of the vocabularies presented in a search history and
Table 4. Comparison result for query 2.

| Evaluation | Dataset Names/Keywords | Precision |
|------------|------------------------|-----------|
| GOOD       | WENTZ_SASS_SIGMA0_L2   | 73.3%     |
|            | OSCAT.LEVEL_2A.V2      |           |
|            | 12 HOURS, OCEAN, DATASET, RADAR, WIND, SEAWINDS, SURFACE, DAY, ICE, VECTOR, HISTORICAL, SPATIAL RESOLUTION, HOUR, QUICKSAT, RAW |
| OK         | RSCAT_L1B.V1.1         | 53.3%     |
|            | RSCAT_L1B.V1.2         |           |
|            | OCEAN, RADAR, SEAWINDS, 12 HOURS, DATASET, VECTOR, WIND, DAY, ANALYSIS, ORBIT, RAW, HDF, QUICKSAT, HOUR, RESOLUTION |
| BAD        | JASON-1.L2_OST_GPS_E  | 46.7%     |
|            | JASON-1.L2.OST_GPR_E_GEODETC |
|            | OCEAN, RADAR, DAY, NETCDF, SURFACE, DATASET, RESOLUTION, WAVE, WIND, ORBIT, WATER, RAW, VECTOR, JASON-1, TOPOGRAPHY |

Table 5. Comparison result for query 3.

| Evaluation  | Dataset Names/Keywords | Precision |
|-------------|------------------------|-----------|
| EXCELLENT   | AVHRR_PATHFINDER_L3_SST_DAILY_DAYTIME_V51 | 80%       |
|            | AVHRR_PATHFINDER_L3_SST_7DAY_DAYTIME_V51 |           |
|            | PATHFINDER, OCEAN, SURFACE, TEMPERATURE, SST, AVHRR-3, DAY, GLOBAL, MONTHLY, HISTORICAL, RAW, NOAA-17, HDF, NOAA-18, DATASET |
| GOOD        | AVHRR_PATHFINDER_L3_SST_ANNUAL_DAYTIME_V51 | 66.7%     |
|            | AVHRR_PATHFINDER_L3_SST_ANNUAL_DAYTIME_V51 |           |
|            | PATHFINDER, OCEAN, SURFACE, TEMPERATURE, SST, AVHRR-3, ANNUAL, NOAA-18, NOAA-17, NOAA-16, MONTHLY, DATASET, HDF, DAY, GLOBAL |
| BAD         | TELLUS_DOT_2008_DATA   | 20%       |
|            | SEAWINDS_BYU_L3_OW_SIGMA0_ARCTIC_POLAR_STEREOGRAPHIC_BROWSE_IMAGES_LTOD |
|            | OCEAN, SURFACE, ASCII, DATASET, DAY, 12 HOURS, SEAWINDS, ADEOS-II, RESOLUTION, QUICKSAT, ICE, HOUR, RADAR, CIRCULATION, TOPOGRAPHY |

describes their relationships based on ontologies and links in a Bayesian Network. Using the Bayesian Network, keywords can be effectively recommended for searches that were done before and for which no search history exists. Essentially, our method bypasses this so-called “cold start” and “new data” problem by building an ontology of concepts and reasoning between them using an ontology-based Bayesian Network model. In general, this model showed significant potential for spatiotemporal data discovery but given its concept focus is not limited to this type of data.

The paper contributes to the following domains. Compared with existing work on recommendation systems such as collaborative filtering, we propose a content-based recommendation method, which resolves the “cold start” and “new data” problem, i.e., infrequent and biased searches, when searching scientific data such as earth observation data. Our model provides a reliable method and it can be seen as an alternative to user-centered recommendations based on specific users’ search behavior and their interactions. From the perspective of knowledge representation and ontologies, this proposed model integrates the quantitative
Bayesian Network model with the conceptual modeling of ontologies. It implements probabilistic reasoning based on ontologies, resulting in improved spatiotemporal data discovery.

Directions for future work are as follows. Firstly, this presented method was mainly focused on the recommendation of keywords. For the recommendation of a specific dataset, more aspects need to be considered. Moreover, this content-based recommendation method could be integrated with collaborative filtering to combine their advantages in recommendation systems. Secondly, this proposed model can be also applied in other fields, such as social media data analysis to improve the functions of keyword analysis in Twitter data. Third, this model could also be integrated with other machine learning methods in recommendation systems to further improve data discovery.

Data availability statement
The data referred to in this paper is not publicly available at the current time.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
Publication of this article was funded in part by the George Mason University Libraries Open Access Publishing Fund.

References
Aggarwal, C. C. (2016). Recommender systems (pp. 1–28). Cham: Springer International Publishing.
Asanov, D. (2011). Algorithms and methods in recommender systems. Berlin, Germany: Berlin Institute of Technology.
Bernard, L., Kanellopoulos, I., Annoni, A., & Smits, P. (2005). The European geoportal—One step towards the establishment of a European spatial data infrastructure. Computers, Environment and Urban Systems, 29(1), 15–31.
Bogdanović, M., Stanimirović, A., & Stoimenov, L. (2015). Methodology for geospatial data source discovery in ontology-driven geo-information integration architectures. Web Semantics: Science, Services and Agents on the World Wide Web, 32, 1–15.
Carvalho, R. N., Laskey, K. B., & Costa, P. C. (2010, November). PR-OWL 2.0-bridging the gap to OWL semantics. In Proceedings of the 6th International Conference on Uncertainty Reasoning for the Semantic Web-Volume 654 (pp. 73–84). Shanghai, China. CEUR-WS.org.
Cui, K., & Bittner, T. (2016). Ontology-based geo-spatial knowledge reasoning system. In H. Onsrud, & W. Kuhn (Eds.), Advancing geographic information science: the past and next twenty years (p. 265). GSDI Association Press.
Da Costa, P. C. G., Laskey, K. B., & Laskey, K. J. (2008). Pr-owl: A bayesian ontology language for the semantic web. In Uncertainty reasoning for the semantic web I (pp. 88–107). Berlin: Springer Berlin Heidelberg.
Edberg, S. J., Evans, D. L., Graf, J. E., Hyon, J. J., Rosen, P. A., & Waliser, D. E. (2016). Studying earth in the new millennium: NASA jet propulsion laboratory’s contributions to earth science and applications space agencies. IEEE Geoscience and Remote Sensing Magazine, 4(1), 26–39.
Giuliani, G., Ray, N., Schwarzer, S., De Bono, A., Peduzzi, P., Dao, H., ... Lehmann, A. (2013). Sharing environmental data through GEOSS. In D. P. Albert and G. R. Dobbs (Eds.), Emerging methods and multidisciplinary applications in geospatial research (pp. 266–281). IGI Global.

Hendler, J., Holm, J., Musialek, C., & Thomas, G. (2012). US government linked open data: Semantic data. *IEEE Intelligent Systems*, 27(3), 25–31.

Hua, X. S., Yang, L., Wang, J., Wang, J., Ye, M., Wang, K., ... Li, J. (2013, October). Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 243–252). Barcelona, Spain: ACM.

Janowicz, K., Schade, S., Bröring, A., Keßler, C., Maué, P., & Stasch, C. (2010). Semantic enablement for spatial data infrastructures. *Transactions in GIS*, 14(2), 111–129.

Jensen, F. V. (1996). *An introduction to Bayesian networks* (Vol. 210, pp. 1–178). London: UCL press.

Jiang, Y., Li, Y., Yang, C., Liu, K., Armstrong, E. M., Huang, T., ... Finch, C. J. (2017). A comprehensive methodology for discovering semantic relationships among geospatial vocabularies using oceanographic data discovery as an example. *International Journal of Geographical Information Science*, 31(11), 2310–2328.

Kourtesis, D., Alvarez-Rodríguez, J. M., & Paraskakis, I. (2014). Semantic-based QoS management in cloud systems: Current status and future challenges. *Future Generation Computer Systems*, 32, 307–323.

Lacasta, J., Nogueras-Isa, J., Béjar, R., Muro-Medrano, P. R., & Zarazaga-Soria, F. J. (2007). A Web ontology service to facilitate interoperability within a spatial data infrastructure: Applicability to discovery. *Data & Knowledge Engineering*, 63(3), 947–971.

Lee, J. G., & Kang, M. (2015). Geospatial big data: Challenges and opportunities. *Big Data Research*, 2(2), 74–81.

Li, W., Yang, C., Nebert, D., Raskin, R., Houser, P., Wu, H., & Li, Z. (2011). Semantic-based web service discovery and chaining for building an Arctic spatial data infrastructure. *Computers & Geosciences*, 37(11), 1752–1762.

Liu, T. Y. (2011). *Learning to rank for information retrieval*. New York, NY: Springer Science & Business Media.

Lutz, M., & Kolas, D. (2007). Rule based discovery in spatial data infrastructure. *Transactions in GIS*, 11(3), 317–336.

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2), 72–79.

Maguire, D. J., & Longley, P. A. (2005). The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems*, 29(1), 3–14.

Mangold, C. (2007). A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies*, 2(1), 23–34.

McCandless, M., Hatcher, E., & Gospodnetic, O. (2010). *Lucene in action: Covers apache lucene 3.0*. Shelter Island, NY: Manning Publications Co.

Middleton, S. E., De Roure, D. C., & Shadbolt, N. R. (2001, October). Capturing knowledge of user preferences: Ontologies in recommender systems. In *Proceedings of the 1st international conference on Knowledge capture* (pp. 100–107). Victoria, BC: ACM.

Pfoser, D., Pitoura, E., & Tryfona, N. (2002, November). Metadata modeling in a global computing environment. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems* (pp. 68–73). McLean, VA: ACM.

Smith, B. (2003). Ontology. In L. Floridi (Eds.), *Blackwell guide to the philosophy of computing and information* (pp. 155–166). Oxford: Blackwell.

Tait, M. G. (2005). Implementing geoportals: Applications of distributed GIS. *Computers, Environment and Urban Systems*, 29(1), 33–47.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth*, 10(1), 13–53.

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems*, 34(4), 264–277.