Semantic Data Mining: A Survey of Ontology-based Approaches

Dejing Dou
Computer and Information Science
University of Oregon
Eugene, OR 97403, USA
Email: dou@cs.uoregon.edu

Hao Wang
Computer and Information Science
University of Oregon
Eugene, OR 97403, USA
Email: csehao@cs.uoregon.edu

Haishan Liu
Computer and Information Science
University of Oregon
Eugene, OR 97403, USA
Email: ahoyleo@cs.uoregon.edu

***Abstract*—*Semantic Data Mining* refers to the data mining tasks that systematically incorporate domain knowledge, especially formal semantics, into the process. In the past, many research efforts have attested the benefits of incorporating domain knowledge in data mining. At the same time, the proliferation of knowledge engineering has enriched the family of domain knowledge, especially formal semantics and Semantic Web ontologies. Ontology is an explicit specification of conceptualization and a formal way to define the semantics of knowledge and data. The formal structure of ontology makes it a nature way to encode domain knowledge for the data mining use. In this survey paper, we introduce general concepts of semantic data mining. We investigate why ontology has the potential to help semantic data mining and how formal semantics in ontologies can be incorporated into the data mining process. We provide detail discussions for the advances and state of art of ontology-based approaches and an introduction of approaches that are based on other form of knowledge representations.**

I. Introduction

Data mining, also known as knowledge discovery from database (KDD), is the process of nontrivial extraction of implicit, previously unknown, and potentially useful information from data [30]. In the past few decades, advances in data mining techniques lead to many remarkable revolutions in data analytics and big data. Data mining also combines techniques from statistics, artificial intelligence, machine learning, database system, and many other disciplines to analyze large data sets. *Semantic Data Mining* refers to data mining tasks that systematically incorporate domain knowledge, especially formal semantics, into the process. The effectiveness of domain knowledge in data mining has been attested in past research efforts. Fayyad et al. [21] claimed that domain knowledge can play an important role in all stages of data mining including, data transformation, feature reduction, algorithm selection, post-processing, model interpretation and so forth. Russell and Norvig [64] believed that an intelligent agent (e.g., a data mining system) must have the ability to obtain the background knowledge and should learn knowledge more effectively with the background knowledge.

Previous semantic data mining research has attested the positive influence of domain knowledge on data mining. For example, the preprocessing can benefit from domain knowledge that can help filter out the redundant or inconsistent data [41], [59]. During the searching and pattern generating process, domain knowledge can work as a set of prior knowledge of constraints to help reduce search space and guide the search path [8], [9]. Further more, the discovered patterns can be cleaned out [49], [48] or made more visible by encoding them in the formal structure of knowledge engineering [76].

To make use of domain knowledge in the data mining process, the first step must account for representing and building the knowledge by models that the computer can further access and process. The proliferation of knowledge engineering (KE) has remarkably enriched the family of domain knowledge with techniques that build and use domain knowledge in a formal way [64]. Ontology is one of successful knowledge engineering advances, which is the explicit specification of a conceptualization [26], [67]. Normally, an ontology is developed to specify a particular domain (e.g., genetics). Such an ontology, often known as a domain ontology, formally specifies the concepts and relationships in that domain. The encoded formal semantics in ontologies is primarily used for effectively sharing and reusing of knowledge and data. Prominent examples of domain ontologies include the Gene Ontology (GO [73]), Unified Medical Language System (UMLS [45]), and more than 300 ontologies in the National Center for Biomedical Ontology (NCBO [2]).

Research in the area of the Semantic Web [10] has led to quite mature standards for modeling and codifying domain knowledge. Today, Semantic Web ontologies become a key technology for intelligent knowledge processing, providing a framework for sharing conceptual models about a domain. The Web Ontology Language (OWL) [1], which has emerged as the de facto standard for defining Semantic Web ontologies, is widely used for this purpose. The Semantic Web technologies that formally represent domain knowledge including structured collection of prior information, inference rules, knowledge enriched datasets etc., could thus develop frameworks for systematic incorporation of domain knowledge in an intelligent data mining environment.

In this survey paper, we study the advances and state of art of semantic data mining. We specifically focus on the ontology-based approaches. The ontology-based approaches for semantic data mining attempt to make use of formal ontologies in the data mining process. This is generally achieved by using the formal definition of concepts and relationships in ontologies as auxiliary information or constraint conditions to guide the data mining process. For example, in classification, ontology can specify the consistency relationships of the classification task. By ruling out the inconsistent search space, the classification task would result in a better accuracy [8]. Further more, structured organization of ontologies can serve as a good representation for the data mining result. For example, in information extraction and text mining, the extracted information can be presented through the ontology itself using an ontology definition language (e.g., OWL) [77]. In this paper, we focus on three perspectives of ontology-based approaches in the research of semantic data mining:

- *Role of ontologies*: Why domain knowledge with formal semantics, such as ontologies, are necessary in all stages of the data mining process.
- *Mining with ontologies*: How ontologies are represented and processed to help the data mining process.
- *Performance evaluation*: How ontologies can improve the performance of data mining systems in applications.

II. Role of Ontologies in Semantic Data Mining

The perspective and mechanism of utilizing ontologies in semantic data mining varies across different systems and applications. The question why ontology is useful in assisting data mining process does not have an uniform conclusion. By reviewing the previous ontology-based approaches, we summarize the following three purposes that ontologies have been introduced to semantic data mining:

- To bridge the semantic gap between the data, applications, data mining algorithms, and data mining results.
- To provide data mining algorithms with a priori knowledge which either guides the mining process or reduces/constrains the search space.
- To provide a formal way for representing the data mining flow, from data preprocessing to mining results.

A. Bridging the semantic gap

The question why domain knowledge is helpful in the data mining process has been long discussed in previous semantic data mining research. Researchers claim that there exists a knowledge gap between the data, data mining algorithm, and mining results in all stages of data mining including preprocessing, algorithm execution, and result generation [18].

Data preprocessing usually contains data cleaning, normalization, transformation, feature extraction and selection. In most scenarios, there exist semantic gaps in the steps of data preprocessing. Without considering formal semantics, ad-hoc or empirical methods are used to determine the quality of the data. For example, scarcity and nearest neighbor rules are usually adopt to determine the outliers and missing values. In the normalization and transformation step, data semantics is necessary for understanding the relations of the data. For example, it is important to determine the correlation between features and attributes of the data when performing data normalization. Strongly correlated attributes could be reduced into one combined attribute. In practice, semantic gaps are usually filled manually by domain experts. However, ontologies have been shown to be beneficial in many data preprocessing tasks [41], [59], [72].

There exists semantic gap between the data mining algorithm and data as well. Data mining algorithms are usually designed for data collected from different domains and scenarios. However, data from a specific domain usually carry domain specific semantics. The generic data mining algorithms lack the ability to identify and make use of semantics across different domains and applications. Ontologies are useful to specify domain semantics and can reduce the semantic gap by annotating the data with rich semantics. Semantic annotation aims at assigning the basic element of information links to formal semantic descriptions [20], [42]. Such elements should constitute the semantics of their source. Semantic annotation is crucial in realizing semantic data mining by bringing formal semantics to data. The annotated data are very convenient for the later steps of semantic data mining because the data are promoted to the formal and structured format that connects ontological terms and relations.

Many research efforts have dedicated to bridge the semantic gap between data mining results and users. Marinica et al. [49], [50] used ontology for the post pruning and filtering of the association rule mining results. Mansingh et al. [48] used ontology to assist the subjective analysis for the association rule post-pruning task. The data mining results can be represented by ontologies in the semantic rich format which help sharing and reuse. For example, information extraction (IE) is the task of automatically extracting structured information from text. The data/text mining results are sets of structured information and knowledge with regarding to the domain. To represent the structured and machine-readable information, it is nature to represent the information with ontology. Ontology Based Information Extraction (OBIE) [77] has extensively used this representation. With OBIE, the information extracted is not only well structured but also represented by predicates in the ontology which are easy for sharing and reuse.

B. Providing prior knowledge and constraints

The definition and reuse of prior knowledge is one of the most important problems for semantic data mining. As a formal specification of concepts and relationships, ontology is a nature way to encode the formal semantics of prior knowledge. The encoded prior knowledge has the potential to guide and influence all stages of the data mining process, from preprocessing to result filtering and representation. For example, Liu et al. [46] developed a RDF hypergraph representation to capture information from both ontologies and data. Ontologies are incorporated into the graph representation of the data as the priori knowledge to bias the graph structure and also representing the distances between terms and concepts in the graph. The approach transforms the hypergraph and weighted hyperedges into a bipartite graph to represent both the data and ontology in a uniformed structure. Random walks with restart over the bipartite graph is performed to generate semantic associations. Whenever the random walk goes through the ontology-based edges, the domain knowledge encoded in ontologies bridges the latent semantic relations underneath the data with rich semantics.

As a collection of concepts and predicates, ontology has the ability to perform logic reasoning and thus make consistency inference for those predicates. In semantic data mining, the ability to make consistency inference is usually represented as constraints. The set of constraints powered by the ontology have the ability to detect inconsistent data and results in the preprocessing stage, the algorithm execution stage, and the result filtering and generation stage. For example, Balcan et al. [8] incorporated ontology as consistency constraints into multiple related classification tasks. The ontology specifies the constraints between multiple classification tasks. Carlson et al. [16] presented a semi-supervised information extraction algorithm that couples the training of many information extractors. Using ontology as constraints on the set of extractors, it yields more accurate results. Claudia Marinica et al. [49], [50] presented post-processing of the association rule mining results using ontology for consistency checking. Invalid or inconsistent association rules are pruned and filtered out with the help of ontology and an inference engine.

C. *Formally representing data mining results*

The well designed data mining systems should present results and discovered patterns in a formal and structured format, so that data mining results are capable to be interpreted as domain knowledge and to further enrich and improve current knowledge bases. Ontology is one of the way to represent the data mining results in a formal and structured way. As a formal definition of concepts and relationships, ontology can encode rich semantics for different domains. The data mining results from different domains and tasks conform naturally with the representation of ontology, for example, information extraction and association rule mining. Specifically, in ontology-based information extraction (OBIE) [55], [77], the extracted information are a set of annotated terms from the document with the relations defined in the ontology. It is therefore straight forward to represent the extracted information with ontology.

Wimalasuriya and Dou [77] claimed that ontology is a valid form to represent the OBIE results in a semantic rich format. Encoding OBIE results in the formal structure of ontology could streamline the data mining process of other data mining tasks that need to make use of the current result. The inference engines which was designed in the field of knowledge engineering could perform consistency checking that validate the data mining results and clean out the inconsistent results. OBIE systems can extract information with higher recall and accuracy compared with traditional IE systems. The ontology in OBIE provides the function as a conceptual framework and consistency checking. It also organizes the extracted information in a formal and structured way using explicit ontology representation. Similarly, ontology-based association pattern mining method [46] can represent latent semantic associations.

III. MINING WITH ONTOLOGIES

With formally encoded semantics, ontology has the potential to assist in various data mining tasks. In this section, we summarize semantic data mining algorithms designed in several important tasks, including association rule mining, classification, clustering, recommendation, information extraction, and link prediction.

A. *Ontology-based Association Rule Mining*

Association rule mining is a fundamental data mining task and well used in different applications. In the early work, Svatek and Rauch [71] designed association mining tool that can benefit from ontologies in all four stages of the mining process: data understanding, task design, result interpretation, and result dissemination over the Semantic Web. Bellandi et al. [9] presented an ontology-based association rule mining method, which queries the ontology to filter the instances used in the association rule mining process. Ontology in this work provides the constraints for queries in the association mining process. The search space of association mining is constrained by the query returned from the ontology that some items from the output association rules are excluded or to be used to characterize interesting items according to an abstraction level. The user constraints include both pruning constraints, which are used for filtering a set of non-interesting items, and abstraction constraints, which permit a generalization of an item to a concept of the ontology.

Marinica et al. [49], [50] presented post-processing of the association rule mining results using an ontology for the consistency checking. Invalid or inconsistent association rules are pruned and filtered out with the help of ontology and an inference engine. Recently, Liu et al. [46] proposed to apply ontology and hypergraph to discover latent association rules in the data. They built the connections between ontology and data using a bipartite hypergraph model. Random walk based metrics were proposed to measure the latent semantic distances between concepts and terms. The term sets with shorter semantic distances are ranked higher. Top ranked term sets are generated as strong associations.

B. *Ontology-based Classification*

Classification is one of the most common data mining tasks that finding a model (or function) to describe and distinguish data classes or concepts [30]. In semantic data mining, one typical use of ontology is to annotate the classification labels with the set of relations defined in the ontology. Research by Balcan et al. [8] indicates that with the ontology annotated classification labels, the semantics encoded in the classification task has the potential not only to influence the labeled data in the classification task but also to handle large number of unlabeled data. They incorporated ontology as consistency constraints into multiple related classification tasks. These tasks classify multiple categories in parallel. An ontology specifies the constraints between the multiple classification tasks. An unlabeled error rate is defined as the probability the classifier assigns a label for the unlabeled data that violates the ontology. This classification task produces the classification hypothesis with the classifiers that produce the least unlabeled error rate and thus most classification consistency.

Allahyari et al. [7] presented an ontology-based method for automatic classification of text documents into a dynamically defined set of topics of interest. Using DBpedia-based ontology, entities and relations among entities are identified from the text document. Semantic graph of connected entities are constructed from the set of relations. HITS algorithm [43] is used to identify the core entities in the semantic graph for the further identification of dynamic topics. The classification of documents is based on calculating the similarity of document's semantic graph to define ontological context (topics).

C. Ontology-based Clustering

Clustering [34] is a data mining task that grouping a set of objects in the same cluster which are similar to each other. Early work of ontology-based clustering includes using ontology in the text clustering task for the data preprocessing [31], enriching term vectors with ontological concepts [32], and promoting distance measure with ontology semantics [36].

Song et al. [65] took advantage of the thesaurus-based and corpus-based ontology for text clustering with the enriched conceptual similarity. They proposed a genetic algorithm for text clustering with transformed latent semantic indexing using ontology to capture the associated semantic similarity. Jing et al. [35] used ontology to re-weight the vectors in knowledge-based vector space for text clustering. Fodeh [23] claimed that ontology can be used to greatly reduce the number of features needed in the document clustering task. With the aid of ontology, a core subset of semantic features for each text corpus is identified. Using this core semantic features for clustering, the number of features can be reduced by 90% or more while still produce clusters that capture the main themes in a text corpus.

Ovaska et al. [58] performed a gene clustering task from microarray experiments with the aid of gene ontology (GO). Graph structure (GS) and information content (IC) based measures are used for the similarity measure between genes. GS-based methods use the hierarchical structure of GO to compute the gene similarity. IC-based methods additionally consider the information content of GO terms in a reference gene set. Zhang et al. [82] proposed medical document clustering with ontology-based term similarity measures. Ontology is used to index the terms in the medical document set. The weight of term is re-calculated by ontology-based term similarity measure. Spherical K-mean is then used for the clustering task.

D. Ontology-based Information Extraction

Information extraction (IE) refers to the task of retrieving certain types of information from natural language text by processing them automatically. IE is closely related to text mining. Ontology-based information extraction (OBIE) is a subfield of information extraction, which uses formal ontologies to guide the extraction process [40], [77]. Because of this guidance in the extraction process, OBIE systems have mostly implemented following a supervised approach [76]. Although very few semi-supervised IE systems are considered as ontology-based [78], [79], they rely on instances of known relationships [4], [63]. Therefore those semi-supervised systems can also be considered as OBIE systems.

Early work of OBIE includes knowledge extraction from web documents [5] and data-rich unstructured documents [19]. Ontology can provide consistency checking for the extracted information in the IE system. Kara [39] presented an ontology-based information extraction and retrieval system which uses ontology for consistency checking. The output of a regular IE system is transformed to ontological instances through ontology population. The inference and consistency checking are performed on these ontological instances. Carlson et al. [17] proposed the semi-supervised information extraction algorithm with few labeled data and large amount of unlabeled data. The proposed algorithm couples multiple ontology based information extractors with ontology specify the constraints and exclusions for different categories and relations. The algorithm iteratively and incrementally enrich the classification label using most confident outputs of these extractors.

Recently, Fernández et al. [22] presented a way to exploit domain knowledge bases to support semantic search capabilities in large document repositories. Nebhi [56] proposed an OBIE system for disambiguating Twitter messages. By combing concepts from Freebase and extraction rules based on dependency trees, Nebhi's approach determines the meaning (and context) of entities mentioned in the messages. Nebhi [57] improved the accuracy the disambiguation process by replacing the pattern-based approach with a classification task, using Support Vector Machine. As a way to promote the adoption of OBIE, Wimalasuriya and Dou [76] proposed the Ontology-based Components for Information Extraction (OBCIE) architecture. OBCIE aims to encourage re-usability by modeling the components of the IE system as modular as possible. Gutierrez et al. [27] extended the OBCIE architecture by incorporating hybrid configurations (e.g., different implementations and different functionalities).

E. Ontology-based Recommendation System

Recommender systems or recommendation systems [3], [12] are the systems that dedicate to predict the preference or ratings that a user would give to an item. Recommendation systems have become extremely popular in recent years and been applied in a variety of applications including movies, music, news, books, research articles, search queries, and social tags [44], [75]. In a good recommendation system, heterogeneous information from multiple sources is usually required. Ontology can integrate the use of heterogeneous information and guide the recommendation preference.

Early work of ontology-based recommendation system uses ontology for user profiling [53], personalized search [60], and web browsing [52], [51]. Recently, Pudota et al. [61] proposed a recommendation system that generate and recommend tags automatically for web resources. The web documents are annotated and matched by terms in the ontology first.

Then ontology-based reasoning is conducted to infer the new knowledge from the annotated terms. This inference is made by finding the common ancestor nodes for them and possibly all the nodes in the path between the matched nodes with ontological concepts. Kang and Choi [38] proposed an ontology-based recommendation system in which the ontology is used to encode the long term and short term preference information. The user preference ontology is constructed from the concepts of the general domain ontology together with the documents that the user visited. Recommendation is made based on the similarity between ontological concepts and terms.

IJntema [33] developed a recommendation system, Athena, to provide ontology-based recommendation for the news feed system. It extend the Herme framekwork [24], a framework used to build a news personalization service, with the help of ontology to determine the semantic relations between terms and concepts. It uses an ontology to store concepts and their relationships to the news items. Cantador et al. [14] proposed another news recommendation system that makes use of ontologies to provide online news recommendation services. Domain ontologies are used to provide the concept framework for news contents and user preferences. Domain ontologies can automatically annotate news items with semantic concepts that appear in both the textual contents and the domain ontologies.

F. Ontology-based Link Prediction

Link prediction for social networks becomes a very active research area in data mining due to the success of online social networks such as Twitter, Facebook, and Google+. Aljandal et al. [6] presented a link prediction framework with ontology-enriched numerical graph features. The authors claimed that in previous social network research flat representation of interest taxonomies limited the improvement of link prediction. Ontology aggregated distance measure is proposed to encode the interest taxonomies in ontology into the distance measure to more accurately describe the shared user interests.

Thor et al. [74] presented a link prediction method on ontology annotated data. The data are first annotated by controlled vocabulary terms from ontologies. The annotation links between the data and predicates in ontology form an annotation graph. Graph summarization and dense subgraph method were proposed to filter the graph and find promising subgraphs. A scoring function based on multiple heuristics was proposed to rank the predictions based on these filtered subgraphs. Amakrishnan [62] proposed a method to discover the informative connection subgraphs that relate two entities in the graph. They proposed heuristics for edge weighting that depend indirectly on the semantics of entity and property types in the ontology and on characteristics of the instance data. The *display ρ-graph generation* algorithm was proposed to extract a small connection subgraph from the input graph. Mabroukeh and Ezeife [47] proposed using the domain ontology for semantic web usage mining and next page prediction. Semantic information in the ontology is used in the sequential pattern mining algorithm to prune the search space and partially relieve the algorithm from support counting.

IV. Performance Evaluation and Applications

As a formal specification of domain concepts and relationships, ontology can assist in the data mining process in various perspectives. It is reasonable to expect a performance gain in ontology-based approaches compared with the data mining approaches without using ontologies or other form of domain knowledge. Many semantic data mining research efforts have attested such improvements. With well designed algorithms, previous research either reports performance improvement or accomplishment of data mining tasks that could not be achieved without using ontologies. In this section, we give a brief summarization of the performance improvement in ontology-based approaches and their applications.

A. Performance gain in precision, recall, and consistency of data mining results

Many previous ontology-based efforts have reported performance gain in the data mining results. Ontology-based approaches have been reported to have better precision and recall than the traditional approaches in various tasks such as text clustering [32], [33], [35], [65], [82], information extraction [17], [27], [56], [57], link prediction [6], [15], [74], and recommendation systems [33], [52], [60], [61].

Research in recommendation system suggests that ontology-based recommendation systems have better prediction precision than traditional recommendation methods [13], [83]. With the enriched semantics and reduced search space, execution speed gain has been reported in the gene clustering task from microarray experiments with ontology-based clustering [58]. In the web usage mining and next page prediction task, semantics-aware sequential pattern mining algorithms was proved to perform 4 times faster than regular and non-semantics-aware algorithms [47].

Ontology-based approaches improve the consistency of data mining results as well. Marinica et al. [49], [50] presented post-processing of the association rule mining results using an ontology for the consistency checking. Semantically inconsistent association rules are pruned and filtered out with the help of ontology and logic reasoning.

B. Semantics rich data mining results

Ontology can also assist in enriching data mining results with formal semantics. Semantics rich data mining results are expected from ontology-based approaches compared with approaches without using ontologies. For example, OBIE is able to extract the information with similar or close semantics that does not directly appear in the data source [39].

Without knowing semantics of the attributes or itemsets, association rule mining usually generate too many rules or even inconsistent rules. Ontology-based association rule mining bridges the semantic gap of the domain knowledge and the association rule mining algorithm. It results in better collection and representation of association rules by pruning the results or reducing the search space. The top ranked rules also result in high support measure for the targeting domain [9].

With the aid of ontology, multi-level association rule mining will discover concept-based association rules instead instance-based rules [29], [66], [80]. With supermarket transactions like *cheese* and *milk*, *bread* and *cake*, etc., traditional association rule mining methods have to generate rules with those items, while multi-level association rule can generate conceptual level rules like *diary product* → *bakery products*. The well controlled granularity of semantics raises the potential of more interesting association rules.

C. Performing data mining task that is unachievable with traditional data mining methods

Certain data mining tasks that are not achievable by traditional data mining methods can be accomplished by ontology-based approaches. For example, traditional classification task usually requires at least reasonable amount of labeled data as prior knowledge. Using ontology as the specification of prior knowledge, classification task without enough labeled data is proved to have a comparable performance compared with traditional classification methods [8]. Using the ontology as a conceptual consistency constraint, the model with unlabeled data can be tuned into the one that have the best consistency with the prior knowledge (i.e., ontology). Classification task without labeled or annotated data is also reported in the ontology-based text classification task [7].

V. Other Approaches in Semantic Data Mining

Although ontology is one of the most common ways for formally representing domain knowledge, other representations of domain knowledge have been used in semantic data mining. Early research efforts of semantic data mining have employed concept hierarchy as a very important representation of domain knowledge. Previous concept hierarchy based algorithms largely focus on exploiting its generalization ability that it could handle the raw data at higher conceptual level. Han et al. [28] use concept hierarchy to guide such generalization process of attributes in quantitative association mining. Later, Han and Fu [29] proposed multi-level association rule mining using concept hierarchy to control the granularity of knowledge discovered from data at different conceptual levels. Kamber et al. [37] proposed concept hierarchy based decision tree methods in which the induction of decision trees could be achieved at different levels of conceptual abstraction.

Later, some research efforts have employed knowledge bases for semantic data mining tasks, including Wikipedia and Freebase [11], which are not exactly formal ontologies. Gabrilovich and Markovitch [25] computed the semantic relatedness using Wikipedia-based semantic analysis in which substantial improvements in computing word and text relatedness is confirmed. Milne and Witten [54] deployed Wikipedia as the external knowledge for the document clustering task. Significant performance improvement has been achieved using concept and category information in Wikipedia to annotate the documents with enriched semantics information. Yu et al. [81] explored a way to build personalized entity recommendation framework for search engine users by utilizing the knowledge extracted from Freebase. A user log dataset collected from a commercial search engine together with the entity graph extracted from Freebase are used to generate semantic enriched features and build up recommendation models.

Most recently, a new representation, *Meta-path*, has been designed for semantic data mining tasks. The meta-path is a path that defines a composition of relations between the set of terms on the path [69]. For example, in the bibliographic network, a typical meta-path could be *author* → *paper* → *venue*. It is usually defined based on the graph of network schema of related data mining terms and concepts. Comparing with an ontology, each meta-path could relate to multiple concepts while each predicate in an OWL ontology usually is related to two concepts. The type of meta-path is defined by the type of entities in the meta-path while the type of predicate in an ontology is defined by the related concepts. Recent research efforts on meta-path have successfully shown its capability to explore efficient semantic data mining algorithm from many perspective. Early work of meta-path focuses on exploiting the semantic enriched similarity representation in heterogenous information network on various applications [68], [70]. Sun et al. [68] presented PathSim, a meta-path based similarity search method in which the semantic similarities between entities are calculated according to the structure of meta-path and the matrix representation of data instance relations. Sun et al. [70] also proposed PathSelClus, a meta-path based clustering algorithm in which semantic similarities are measured based on a probabilistic model on the meta-path framework.

VI. Conclusion

The advances in knowledge engineering and data mining promote semantic data mining, which brings rich semantics to all stages of data mining process. Many research efforts have attested the advantage of incorporating domain knowledge into data mining. Formal semantics encoded in the ontology is well structured which is easy for the machine to read and process thus make it a nature way to use ontologies in semantic data mining. Using ontologies, semantic data mining has advantages to bridge semantic gaps between the data, applications, data mining algorithms, and data mining results, provide the data mining algorithm with priori knowledge which either guides the mining process or reduces the search space, and to provide a formal way for representing the data mining flow, from data preprocessing to mining results.

In the past decade, to handle and manipulate the big data have raised intense discussion in the data mining community. With the development of knowledge engineering, especially Semantic Web techniques, mining large amount, semantics rich, and heterogeneous data emerges as an important research topic in the community. As many researchers have pointed out, work along semantic data mining is still in its early stage. Ontology-based semantic data mining seems to be one of most promising approaches. The major challenge is to develop more automatic semantic data mining algorithms and systems by utilizing the full strength of formal ontology that has well defined representation language, formal semantics, and reasoning tools for logic inference and consistency checking.

Acknowledgment

We would like to thank the paper invitation from the organizing committee of the 2015 IEEE International Conference on Semantic Computing. This research has been partially supported by the NSF grant IIS-1118050 and NIH grant R01GM103309. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NSF, the NIH, or the U.S. Government.

References

[1] OWL Web Ontology Language. `http://www.w3.org/TR/owl-ref/`.

[2] The National Center for Biomedical Ontology. `http://www.bioontology.org/`.

[3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[4] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 85–94, 2000.

[5] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt. Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems, IEEE*, 18(1):14–21, 2003.

[6] W. Aljandal, V. Bahirwani, D. Caragea, and W. H. Hsu. Ontology-aware classification and association rule mining for interest and link prediction in social networks. In *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pages 3–8, 2009.

[7] M. Allahyari, K. J. Kochut, and M. Janik. Ontology-based text classification into dynamically defined topics. In *Semantic Computing (ICSC), 2014 IEEE International Conference on*, pages 273–278, 2014.

[8] N. Balcan, A. Blum, and Y. Mansour. Exploiting ontology structures and unlabeled data for learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1112–1120, 2013.

[9] A. Bellandi, B. Furletti, V. Grossi, and A. Romei. Ontology-driven association rule extraction: A case study. *Contexts and Ontologies Representation and Reasoning*, page 10, 2007.

[10] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):28–37, 2001.

[11] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

[12] R. Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[13] I. Cantador, A. Bellogín, and P. Castells. A multilayer ontology-based hybrid recommendation model. *AI Communications*, 21(2):203–210, 2008.

[14] I. Cantador, A. Bellogín, and P. Castells. Ontology-based personalised and context-aware recommendations of news items. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 562–565, 2008.

[15] D. Caragea, V. Bahirwani, W. Aljandal, and W. H. Hsu. Ontology-based link prediction in the livejournal social network. In *SARA*, volume 9, pages 1–1, 2009.

[16] A. Carlson, J. Betteridge, E. R. Hruschka Jr, and T. M. Mitchell. Coupling semi-supervised learning of categories and relations. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 1–9, 2009.

[17] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, and T. M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web Search and Data Mining*, pages 101–110, 2010.

[18] P. Domingos. Toward knowledge-rich data mining. *Data Mining and Knowledge Discovery*, 15(1):21–28, 2007.

[19] D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddle. Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 52–59. ACM, 1998.

[20] M. Erdmann, A. Maedche, H.-P. Schnurr, and S. Staab. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In *Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, pages 79–85, 2000.

[21] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[22] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta. Semantically enhanced information retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):434–452, 2011.

[23] S. Fodeh, B. Punch, and P.-N. Tan. On ontology-driven document clustering using core semantic features. *Knowledge and information systems*, 28(2):395–421, 2011.

[24] F. Frasincar, J. Borsje, and L. Levering. A semantic web-based approach for building personalized news services. *International Journal of E-Business Research (IJEBR)*, 5(3):35–53, 2009.

[25] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

[26] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5):907–928, 1995.

[27] F. Gutierrez, D. Dou, A. Martini, S. Fickas, and H. Zong. Hybrid ontology-based information extraction for automated text grading. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 359–364. IEEE, 2013.

[28] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *Knowledge and Data Engineering, IEEE Transactions on*, 5(1):29–40, 1993.

[29] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *VLDB*, volume 95, pages 420–431, 1995.

[30] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[31] A. Hotho, A. Maedche, and S. Staab. Ontology-based text document clustering. *KI*, 16(4):48–54, 2002.

[32] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Proceedings of the third IEEE International Conference on Data Mining*, pages 541–544, 2003.

[33] W. IJntema, F. Goossen, F. Frasincar, and F. Hogenboom. Ontology-based news recommendation. In *Proceedings of the 2010 EDBT/ICDT Workshops*, page 16. ACM, 2010.

[34] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[35] L. Jing, M. K. Ng, and J. Z. Huang. Knowledge-based vector space model for text clustering. *Knowledge and information systems*, 25(1):35–55, 2010.

[36] L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang. Ontology-based distance measure for text clustering. In *Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA*, 2006.

[37] M. Kamber, L. Winstone, W. Gong, S. Cheng, and J. Han. Generalization and decision tree induction: Efficient classification in data mining. In *Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on*, pages 111–120. IEEE, 1997.

[38] J. Kang and J. Choi. An ontology-based recommendation system using long-term and short-term preferences. In *Information Science and Applications (ICISA), 2011 International Conference on*, pages 1–8. IEEE, 2011.

[39] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan. An ontology-based retrieval system using semantic indexing. *Information Systems*, 37(4):294–305, 2012.

[40] V. Karkaletsis, P. Fragkou, G. Petasis, and E. Iosif. Ontology based information extraction from text. In G. Paliouras, C. Spyropoulos, and G. Tsatsaronis, editors, *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, volume 6050 of *Lecture Notes in Computer Science*, pages 89–109. Springer Berlin Heidelberg, 2011.

[41] N. Khasawneh and C.-C. Chan. Active user-based and ontology-based web log data preprocessing for web usage mining. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 325–328, 2006.

[42] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[43] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: Measurements, models, and methods. In *Computing and combinatorics*, pages 1–17. Springer, 1999.

[44] Y. Koren. The bellkor solution to the netflix grand prize. *Netflix prize documentation*, 81, 2009.

[45] D. Lindberg, B. Humphries, and A. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291, 1993.

[46] H. Liu, D. Dou, R. Jin, P. LePendu, and N. Shah. Mining biomedical ontologies and data using rdf hypergraphs. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 141–146. IEEE, 2013.

[47] N. R. Mabroukeh and C. I. Ezeife. Using domain ontology for semantic web usage mining and next page prediction. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1677–1680. ACM, 2009.

[48] G. Mansingh, K.-M. Osei-Bryson, and H. Reichgelt. Using ontologies to facilitate post-processing of association rules by domain experts. *Information Sciences*, 181(3):419–434, 2011.

[49] C. Marinica and F. Guillet. Knowledge-based interactive postmining of association rules using ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, 22(6):784–797, 2010.

[50] C. Marinica, F. Guillet, and H. Briand. Post-processing of discovered association rules using ontologies. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*, pages 126–133, 2008.

[51] S. E. Middleton, H. Alani, and D. C. De Roure. Exploiting synergy between ontologies and recommender systems. *arXiv preprint cs/0204012*, 2002.

[52] S. E. Middleton, D. C. De Roure, and N. R. Shadbolt. Capturing knowledge of user preferences: ontologies in recommender systems. In *Proceedings of the 1st international conference on Knowledge capture*, pages 100–107. ACM, 2001.

[53] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88, 2004.

[54] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30, 2008.

[55] H.-M. Müller, E. E. Kenny, and P. W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, 2(11):e309, 2004.

[56] K. Nebhi. Ontology-based information extraction from twitter. In *Workshop on Information Extraction and Entity Analytics on Social Media Data - COLING 2012*, pages 17–22, 2012.

[57] K. Nebhi. Named entity disambiguation using freebase and syntactic parsing. In *LD4IE@ISWC*, 2013.

[58] K. Ovaska, M. Laakso, and S. Hautaniemi. Fast gene ontology based clustering for microarray experiments. *BioData mining*, 1(1):11, 2008.

[59] D. Perez-Rey, A. Anguita, and J. Crespo. Ontodataclean: Ontology-based integration and preprocessing of distributed data. In *Biological and Medical Data Analysis*, pages 262–272. Springer, 2006.

[60] A. Pretschner and S. Gauch. Ontology based personalized search. In *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, pages 391–398, 1999.

[61] N. Pudota, A. Dattolo, A. Baruzzo, F. Ferrara, and C. Tasso. Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems*, 25(12):1158–1186, 2010.

[62] C. Ramakrishnan, W. H. Milnor, M. Perry, and A. P. Sheth. Discovering informative connection subgraphs in multi-relational graphs. *ACM SIGKDD Explorations Newsletter*, 7(2):56–63, 2005.

[63] S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, ECML PKDD'10, pages 148–163, Berlin, Heidelberg, 2010. Springer-Verlag.

[64] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003.

[65] W. Song, C. H. Li, and S. C. Park. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. *Expert Systems with Applications*, 36(5):9095–9104, 2009.

[66] R. Srikant and R. Agrawal. Mining generalized association rules. In *VLDB*, volume 95, pages 407–419, 1995.

[67] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1):161–197, 1998.

[68] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *VLDB'11*, pages 992–1003, 2011.

[69] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1348–1356, 2012.

[70] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):11, 2013.

[71] V. Svátek, J. Rauch, and M. Ralbovskỳ. Ontology-enhanced association mining. In *Semantics, Web and Mining*, pages 163–179. Springer, 2006.

[72] D. Tanasa and B. Trousse. Advanced data preprocessing for intersites web usage mining. *Intelligent Systems, IEEE*, 19(2):59–65, 2004.

[73] The_gene_ontology_consortium. Creating the gene ontology resource: design and implementation. *Genome Res.*, 11(8):1425–1433, August 2001.

[74] A. Thor, P. Anderson, L. Raschid, S. Navlakha, B. Saha, S. Khuller, and X.-N. Zhang. Link prediction for annotation graphs using graph summarization. In *The Semantic Web–ISWC 2011*, pages 714–729. Springer, 2011.

[75] A. Töscher, M. Jahrer, and R. M. Bell. The bigchaos solution to the netflix grand prize. *Netflix prize documentation*, 2009.

[76] D. C. Wimalasuriya and D. Dou. Components for information extraction: Ontology-based information extractors and generic platforms. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM)*, pages 9–18, 2010.

[77] D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323, 2010.

[78] F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, CIKM '07, pages 41–50, 2007.

[79] F. Wu and D. S. Weld. Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 635–644. ACM, 2008.

[80] G. Yang, K. Shimada, S. Mabu, K. Hirasawa, and J. Hu. A genetic network programming based method to mine generalized association rules with ontology. In *SICE, 2007 Annual Conference*, pages 2715–2722. IEEE, 2007.

[81] X. Yu, H. Ma, B.-J. P. Hsu, and J. Han. On building entity recommender systems using user click log and freebase knowledge. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 263–272, 2014.

[82] X. Zhang, L. Jing, X. Hu, M. Ng, J. Xia, and X. Zhou. Medical document clustering using ontology-based term similarity measures. *International Journal of Data Warehousing and Mining (IJDWM)*, 4(1):62–73, 2008.

[83] L. Zhuhadar, O. Nasraoui, R. Wyatt, and E. Romero. Multi-model ontology-based hybrid recommender system in e-learning domain. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03*, pages 91–95, 2009.