ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

Ontology based semantic aware Bayesian Network model

Ramya R, K Pramilarani, Sheba Pari N

P.G. Student, Department of CSE, New Horizon College Of Engineering, Bangalore, Karnataka, India
Senior Assistant Professor, Department of CSE, New Horizon College Of Engineering, Bangalore, ,India
Assistant Professor, Department of CSE, New Horizon College Of Engineering, Bangalore, Karnataka ,India

ABSTRACT: The paper presents a semantic annotation framework that is capable of extracting relevant information from unstructured, ungrammatical and incoherent data sources. The framework uses ontology to conceptualize a problem domain and to extract data from the given corpora, and Bayesian networks to resolve conflicts and to predict missing data. The framework is extensible as it is capable of dynamically extracting data from any problem domain given a predefined ontology and a corresponding Bayesian network. Uncertainty in context awareness always exists in any context-awareness computing. This will lead to imperfectness and incompleteness of sensed data, because of this reason, we must improve the accuracy of context awareness. Proposed system uses ontology and context reasoning method based on Bayesian Network for handling the uncertainty

KEYWORDS: Ontology, Bayesian network, Data mining, Semantic Image

I. INTRODUCTION

Data mining, otherwise called knowledge discovery from database (KDD), is the methodology of nontrivial extraction of verifiable, beforehand obscure, and possibly valuable data from information. With the recent advances in data mining strategies lead to numerous momentous upsets in information investigation and big data. Data mining likewise joins methods from insights, computerized reasoning, machine learning, database framework, and numerous different controls to examine substantial information sets. Semantic Data Mining alludes to data mining errands that deliberately fuse space learning, particularly formal semantics, into the methodology. Past semantic data mining exploration has witness to the positive impact of space learning on data mining. Amid the seeking and pattern generating procedure, area learning can function as an arrangement of former information of requirements to help decrease hunt space and aide the inquiry way

A large volume of web contents is available in unstructured and semi-structured format. This includes the contents available on many online ad portals such as craigslist, eBay, gum tree, etc. Despite providing a reliable and affordable service to a large customer/fan base, these portals contain user-generated posts (data) written in unstructured and ungrammatical format. Thus, standard query languages cannot be used to retrieve relevant information posted on these portals. As a result, users have to rely on key-word-based searches which do not necessarily retrieve the most relevant and accurate information.

II. PROBLEM FORMULATION

The system is an Ontology-Based Text-Mining Method to cluster research proposals based on their similarities in research areas. An ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts. It consists of an axioms, relationships and set of concepts that describe a domain of interests and represents an agreed-upon conceptualization of the domain's "real-world" setting. Implicit knowledge for humans is made explicit for computers by ontology. Thus, ontology can automate information processing and can facilitate text mining in a specific domain (such as research project selection). An ontology based text mining framework has been built for clustering the research proposals according to their discipline areas.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

Uncertainty in context awareness always exists in any context-awareness computing. This will lead to imperfectness and incompleteness of sensed data, because of this reason, we must improve the accuracy of context awareness. Proposed system uses ontology and context reasoning method based on Bayesian Network for handling the uncertainty.

III. MINING WITH ONTOLOGIES AND BAYESIAN NETWORK LEARNING

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. The 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 Webpresented 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.

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. 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 indicates that with the ontology annotated classification labels, the semantics encoded in the classification task has the potential not only to influence the labelled 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.

C. Ontology-based Clustering

Clustering 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, enriching term vectors with ontological concepts, and promoting distance measure with ontology semantics.

D. Ontology-based Information Extraction

Information extraction 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 is a subfield of information extraction, which uses formal ontologies to guide the extraction process. Because of this guidance in the extraction process, OBIE systems have mostly implemented following a supervised approach. Although very few semi-

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

supervised IE systems are considered as ontology-based they rely on instances of known relationships. Therefore those semi supervised systems can also be considered as OBIE systems.

E. Bayesian network learning

To extract information from ungrammatical and incoherent data sources, one has to deal with variable size of information. For instance, if we want to search used cell phones on two different web sites, we may find one website containing very detailed information (brand, model, camera, condition, battery, display, weight, etc.) while the other contains very few data elements (brand and model). Most information extraction systems, when confronted with such situation, store the unavailable data items as missing. However, it is possible in many situations that the data is missing due to some causal reason. In other words, the data is missing be-cause the person who entered the information, in case of online ad portals, thought that it could easily be predicted from other non-missing values.
If the end-user of the posted information is also familiar with such assumptions then no ambiguity arises. However, such assumptions are typically not known to naı¨ve users and they intend to ignore records having missing values. Further-more, in many cases, the relationship between missing and non-missing values is not deterministically causal, but is probabilistically causal. It is thus highly desirable in situations like these to predict the missing data to better guide a user in his/her decision making process. Another major limitation of the existing information extraction systems is their inability to resolve synonym and polysemy. In some cases, context words can aid in resolving this is-sue but the situation becomes complicated when relevant context words are not available in the text or same context words are used for different attributes.

III. AHP ALGORITHM

Analytic Hierarchy Process (AHP) is one of Multi Criteria decision making method that was originally developed by Prof. Thomas In short, it is a method to derive ratio scales from paired comparisons. The input can be obtained from actual measurement such as price, weight etc., or from subjective opinion such as satisfaction feelings and preference. AHP allow some small inconsistency in judgment because human is not always consistent. The ratio scales are derived from the principal Eigen vectors and the consistency index is derived from the principal Eigen value.

The most creative task in making a decision is to choose the factors that are important for that decision. In the Analytic Hierarchy Process we arrange these factors, once selected, in a hierarchy structure descending from an overall to criteria, sub criteria and alternative in successive levels.

IV SYSTEM DESIGN AND IMPLEMENTATION

A.INFORMATION EXTRACTION

The essential components extract information from existing web contents are shown in Fig. 5. The proposed system performs information extraction in two phases. In Phase-I, it uses ontology to extract relevant data while in Phase-II, it uses Bayesian Network to select the most appropriate value (if more than one value is extracted) of an attribute or to predict a value (if the value is missing). The following subsections explain each phase in detail.

B. EXTRACTION VIA ONTOLOGY

During the ontology-based extraction phase, it retrieves links of information of interest from explicitly provided URL(s). In an iterative manner, contents of each link are explored to extract relevant data. The system performs extraction using (a) knowledge stored in an ontology in the form of concepts, relationships among concepts, data type properties, and context keywords and (b) generated rules according to the type (int, float, string) of an attribute. Each context keyword is searched within the ad contents. If a match is found, this suggests that the corresponding data value should be present in the neighborhood of this keyword and can be searched using pre-defined rules/patterns. A fixed threshold of 8 characters is used in the proposed system to define the neighborhood. For example, if the word ‘‘GB’’ is

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

matched then 8 characters before and after it are treated as neighborhood and are searched using the regular expression defined for ‘‘harddisk’’ attribute. Besides 8, several other numbers were also tried, but it was found that 8 pro-vides a good trade-off between efficiency and accuracy. The complete ontology-based extraction process is depicted in Fig. 1

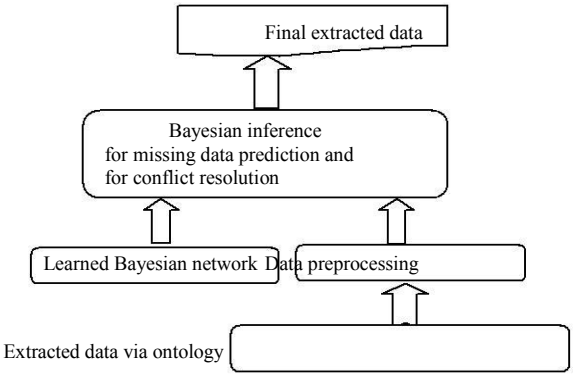

Fig. 1. Phase-I of information extraction.

C .CHALLENGES OF EXTRACTING DATA FROM ONLINE AD PORTALS

Following are the major challenges faced during the application of to extract data from the selected online ad portals.

URL unrecognized: To process the information available on ad portals, the links of all relevant documents have to be retrieved. At many times during this retrieval process, links were found to be removed from the corresponding web sites or were unavailable due to network problems.

Ungrammatical/spelling mistakes: Information posted on online ad portals is not necessarily available in a proper grammatical form. Some ads use abbreviations of different terms and some use different conventions for similar data elements. This leads to higher chances of typing mistakes.

Variable Size of information: Ads’ size varies tremendously depending upon the level of detail provided by different users. Some users provide very detailed information including photo-graphs of the item, while some users simply write a phrase highlighting the most important features of an item.

Appearance: Different data elements with similar appearances (e.g. 10 GB could be related to ram or to hard disk) and same data elements with different appearances lead to identification problems. In car advertisements, mileage appears in different formats such as and sometimes its context words (such as mileage) are not used. This makes it difficult to develop a generalized regular expression as data could belong to integer in one instance and to string in another. It also becomes difficult to find location in the absence of context words.

Unrecognized: Sometimes the required information is available in a unique format which may not be easily recognizable. For instance, the possible values of a string type attribute are stored in the comment section of ontology. Values other than those are considered as unrecognized. Similarly, the neighborhood area of context words is pre-defined and a value that occurs beyond this range is not be recognized.

D. RESULTS OF EXTRACTION VIA ONTOLOGY

The accuracy of ontology based extraction process is measured through recall and precision metrics. The extraction of an attribute can result in three possibilities: value correctly identified (V), missing value correctly identified (M), and value incorrectly identified (W). It should be mentioned that in case of multiple values being extracted, if the correct value exists in the list of values then it is treated as correct extraction (only at this stage). Mathematically.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

V. 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 prior 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.

The paper presented employs ontology and Bayesian networks to extract and annotate data available at unstructured and ungrammatical data sources. The framework is highly extensible as it has the capability to extract data from any problem domain once provided with a domain-specific ontology and the corresponding Bayesian network. It extracts the required information in two phases. The performance of the system was tested on three real data sets available online at the craigslist web-site. The results are very promising with a very high precision and recall ratio and the focus would also be on extracting data from multiple websites belonging to the same problem domain. Handling of spelling errors as well as handling of string data types with large number of possible values are other areas of future research.

REFERENCES

[1] B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, G. Stumme, A roadmap for web mining: from web to Semantic Web, Web Mining: FromWeb to SemanticWeb (2004) 1–22.
[2] B. Berendt, A. Hotho, G. Stumme, Towards Semantic Web mining, in: Proceedings of the First International Semantic Web Conference on The Semantic Web, Springer-Verlag, 2002, pp. 264–278.
[3] A. Tjoa, A. Andjomshoaa, F. Shayeganfar, R. Wagner, Semantic Web challenges and new requirements, in: Sixteenth International Workshop on Database and Expert Systems Applications, 2005, pp. 1160–1163.
[4] M. Wilson, B. Matthews, The Semantic Web: prospects and challenges, in: 7th International Baltic Conference on Databases and Information Systems, 2006, pp. 26–29.
[5] P. Mika, Social Networks and the Semantic Web, Springer, 2007.
[6] G. Antoniou, F.V. Harmelen, A Semantic Web Primer, MIT Press, 2004.
[7] Q.N. Rajput, S. Haider, N. Touheed, Information extraction from unstructured and ungrammatical data sources for semantic annotation, International Journal of Information Technology 5 (3) (2009) 189–197.
[8] Q.N. Rajput, S. Haider, Use of Bayesian network in information extraction from unstructured data sources, International Journal of Information Technology 5
(4) (2009) 207–213.
[9] G. Fiumara, Automated information extraction from Web sources: a survey, in: Proceedings of the Workshop between Ontologies and Folksonomies (BOF), Michigan, USA, 2007.
[10] A.H.F. Laender, B.A. Ribeiro-Neto, A.S.D. Silva, J.S. Teixeira, A brief survey of web data extraction tools, Sigmod Record 31 (2002) 84–93.
[11] L. Reeve, H. Han, Survey of semantic annotation platforms, in: Proceedings of the 2005 ACM Symposium on Applied Computing, Santa Fe, New Mexico, pp. 1634–1638.
[12] J. Tang, J. Li, H. Lu, B. Liang, X. Huang, K. Wang, iASA: learning to annotate the Semantic Web, Journal on Data Semantics IV (2005) 110–145.
[13] M. Motta, E. Motta, J. Domingue, M. Lanzoni, F. Ciravegna, MnM: ontology driven semi-automatic and automatic support for semantic markup, in: Gomez-Perez (Ed.), The 13th International Conference on Knowledge Engineering and Management (EKAW), Springer-Verlag, 2002, pp. 379–391.
[14] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, A. Kirilov, KIM-a semantic platform for information extraction and retrieval, Natural Language Engineering 10 (2004) 375–392.
[15] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y. Ng, R.D. Smith, Conceptual-model-based data extraction from multiple-record Web pages, Data Knowledge Engineering 31 (1999) 227–251.
[16] Y. Ding, D. Embley, S. Liddle, Automatic creation and simplified querying of Semantic Web content: an approach based on information-extraction ontologies, in: The Semantic Web – ASWC 2006, 2006, pp. 400–414.
[17] D.W. Embley, C. Tao, S.W. Liddle, Automating the extraction of data from HTML tables with unknown structure, Data and Knowledge Engineering. 54 (2005) 3–28.
[18] M. Michelson, C.A. Knoblock, Creating relational data from unstructured and ungrammatical data sources, Journal of Artificial Intelligence Research 31 (2008) 543–590.