ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Study of Different Approaches for Information Extraction

Shevali Agarwal
Department of Computer Science, Acropolis Institute of Technology & Research, Indore, India
shevaliagarwal@gmail.com

Abstract

Data mining is the term used to describe the process of extracting valuable information from a database. In this survey paper, we offer a revision of the Data Mining & its Uses and tools. This survey paper is also focused on the techniques of Data Mining & its Algorithm, Application, benefits and drawbacks. The goal of this review is to provide a comprehensive review of Information Extraction using Data Mining Technique.

Keywords: Data mining, Data mining life cycle, Data mining Technique, Data mining applications/Uses, Association Rule, Classification, Clustering, Predication, Decision Tree, Neural Network, Ontology..

Introduction

There is wide growth in the field of World Wide Web. The passive amount of information is available on the internet. Due to the different kind of information structured, unstructured and semi structured and also the lack of structure of the web information sources, it becomes more difficult to retrieve or extract. [1,2] Due to this problem the searching and browsing becomes limited. classy Web mining applications, require expensive maintenance to deal with different data formats. To mechanize the paraphrase of input pages into structured data, plenty of hard work has been devoted in the area of information extraction (IE). Unlike information retrieval (IR), which concerns how to categorize relevant documents from a document collection, IE produces structured data ready for post-processing, which is crucial to many applications of Web mining and searching tools.

We will see different techniques in section 2, in section 3 the conclusion and future enhancement and in section 4 the references.

Techniques for Information Extraction

There are various techniques used to extract the information. In this paper we will study all those techniques used in the field of web mining.

A. Association: Association analysis is considered as a supervised technique. The basic use of this technique is to discover the knowledge discovery task. Association rule mining is also known as association leaning techniques.[3,4,5] It can also find some relationship between the values. Suppose for example, we will consider the supermarket or reliance fresh, here in this the customer purchases the different items. On the regularly purchasing of some items, the supermarket or reliance fresh try to identify the pattern of purchase and on the basis of that they give the different discount to the customer. This is one of the efficient algorithms of data mining.

Apriori is a classic algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

Advantages:

1. Uses large item set property.
2. Easily parallelized
3. Easy to implement.

Disadvantages:

1. Assumes transaction database is memory resident.
2. Requires many database scans.
3. Obtaining non interesting rules
4. Huge number of discovered rules
5. Low algorithm performance

B. Classification: is the data mining techniques based on the machine learning. The purpose of the classification is to classify the each item of the data into predefines set of groups. This is also considered the one of the supervised algorithm. This is very useful for the future prediction. For example certain peoples are also go outside for eating food. Say we take the example of Mac D. on the basis of the customer pattern, the Mac D try to identify the pattern when customer will come next time to Mac

ISSN: 2277-9655
Impact Factor: 1.852

D to order the same thing or we will take example of organization where on the basis of the records of the employee the company try to predict when that person will leave the job.

C. Clustering: is an unsupervised learning algorithm in which objects have somehow of same characteristics can be put in cluster. That is how we can cluster the whole information in different clusters which have the same characteristics. Clustering learning in some way different from the classification based learning. In classification based learning we can put the set of items into predefined classes. But, in clustered based learning we also defined the classes. For Example, suppose in the college there are many faculties who are working. We can distinguish them by making the cluster of the department in the college. Each department is now maintaining the records of those faculties who are working in their department.

Decision Tree: **Decision tree learning** uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. More descriptive names for such tree models are **classification trees** or **regression trees**. In these tree structures, leaves symbolize class labels and branches symbolize conjunctions of features that lead to those class labels.

In decision tree analysis, a decision tree can be used to represent decisions and decision making. In data mining, a decision tree tells about data but it cannot make any decisions; rather the consequential classification tree can be an input for decision making.

Decision trees used in data mining are of two main types:

- **Classification tree** analysis is when the predicted outcome is the class to which the data belongs.
- **Regression tree** analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

D. The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. decision tree used in both the fields regression and classification have some similarities and also having some differences. Such as the way of dividing the tree.

E. Neural Network: is a mathematical or computational model. In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. Neural networks essentially comprise three pieces: the architecture or model; the learning algorithm; and the activation functions. Neural networks are programmed or "trained" to "store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill-defined problems. It is precisely these two abilities (pattern recognition and function estimation) which make artificial neural networks (ANN) so prevalent a utility in data mining. As data sets grow to massive sizes, the need for automated processing becomes clear.

F. Ontology: Ontology-Based Information Extraction (OBIE) has recently emerged as a subfield of information extraction. Here, ontologies are used by the information extraction process and the output is generally presented through ontology. It should be noted that ontology is defined as a formal and explicit specification of a shared conceptualization [6, 7]. Generally, ontologies are specified for particular domains. Since information extraction is essentially concerned with the task of retrieving information for a particular domain, formally and explicitly specifying the concepts of that domain through an ontology can be helpful to this process. For example, a geopolitical ontology that defines concepts like country, province and city can be used to guide the information extraction system described earlier. This is the general idea behind ontology-based information extraction. It appears that the term "Ontology-Based Information Extraction" has been conceived only a few years ago. But there has been some work related to this field before that (e.g., work by Hwang [8] on constructing ontologies from text, published in 1999). Recently, there have been many publications that describe OBIE systems and even a workshop has been organized on this topic [9]. Several of these systems are related to ongoing projects. This, together with the fact that the interest on information extraction in general is on the rise, indicate that this field could experience a significant growth in the near future. Although the field of information extraction appears to be at least few years old it appears that there have not been any serious attempts review the literature of the field and to provide an introduction to the field based on the ideas present in the literature. One aspect of this task is to provide a definition for an OBIE system so that it is possible to clearly identify whether a given systems is an OBIE system or not. In addition, it is necessary to analyze the architectures of different OBIE systems and figure out the commonalities

ISSN: 2277-9655
Impact Factor: 1.852

between them. It is also useful to classify the existing OBIE systems (with respect to suitable dimensions) so that the specialties of different systems can be recognized and new systems can be easily compared against existing ones. Reviewing the implementation details of different OBIE systems and presenting the metrics that researchers have used to evaluate the performance of OBIE systems will also be helpful in developing new OBIE systems.

Conclusion & Future Enhancement

In this paper, we have only study all the important method used for information extractions. We have also see the advantages and disadvantages of different approach. In the next paper we will present the ontology based architecture for information extraction and compare with the existing approach.

References

[1] David Sanchez et al., "Content Annotation for the Semantic Web: An Automatic Web-Based Approach," Knowledge and Information Systems 27 (2011): 393-418.

[2] Seymore et al., "Learning Hidden Markov Model Structure for Information Extraction."

[3] Hipp, J.; Güntzer, U.; Nakhaeizadeh, G. (2000). "Algorithms for association rule mining --- a general survey and comparison". ACM SIGKDD Explorations Newsletter **2**: 58. doi: 10.1145/360402. 360421.

[4] Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" .Introduction to Data Mining. Addison-Wesley. ISBN 0-321-32136-7.

[5] Pei, Jian; Han, Jiawei; and Lakshmanan, Laks V. S.; Mining frequent itemsets with convertible constraints, in Proceedings of the 17th International Conference on Data Engineering, April 2–6, 2001, Heidelberg, Germany, 2001, pages 433-442.

[6] T. R. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition 5(2) (1993) 199-220.

[7] R. Studer, V.R. Benjamins and D. Fensel, Knowledge Engineering: Principles and methods, Data Knowledge Engineering 25(1) (1998) 161-197.

[8] C. Hwang, Incompletely and imprecisely speaking: using dynamic ontologies for representing and retrieving information. In: E. Franconi and M. Kifer (eds), Proceedings of the 6th International Workshop on Knowledge Representation Meets Databases, (ACM, New York, 1999).

[9] B. Adrian, G. Neumann, A. Troussov and B. Popov. In: Proceedings of the First International and KI-08 Workshop on Ontology-Based Information Extraction Systems, (DFKI, Kaiserslautern, Germany, 2008).

[10] J Kleinberg, "An Impossibility Theorem for Clustering", Proceedings of The Neural Information Processing Systems Conference 2002.

[11] D.S. Weld, R. Hoffmann and F. Wu, Using wikipedia to bootstrap open information extraction, SIGMOD Record 37(4) (2008) 62-68.

[12] D.W. Embley, Toward semantic understanding: an approach based on information extraction ontologies. In: Proceedings of the 15th Australasian database conference, (Australian Computer Society, Darlinghurst, Australia, 2004).

[13] A. Maedche, G. Neumann, and S. Staab, Bootstrapping an ontology-based information extraction system. In: P.S. Szczepaniak, J. Segovia, J. Kacprzyk and L.A. Zadeh (eds), Intelligent Exploration of the Web, (Physica-Verlag GmbH, Heidelberg, Germany, 2003).

[14] B. Yildiz and S. Miksch, OntoX - a method for ontology-driven information extraction. In: Proceedings of the 2007 International Conference on Computational Science and Its Applications, (Springer, Berlin, 2007).

[15] P. Buitelaar, P. Cimiano, P. Haase and M. Sintek, Towards linguistically grounded ontologies. In: Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications, (Springer-Verlag, Berlin, 2009).

[16] L. McDowell and M. J. Cafarella, Ontology-driven information extraction with OntoSyphon. In: Proceedings of the 5th International Semantic Web Conference, (Springer, Berlin, 2006).

[17] F. Wu, R. Hoffmann, and D. S. Weld, Information extraction from Wikipedia: moving down the long tail. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (ACM, New York, 2008).

[18] H. Saggion, A. Funk, D. Maynard, and K. Bontcheva, Ontology-based information extraction for business intelligence. In: Proceedings of the 6th International and 2nd

ISSN: 2277-9655
Impact Factor: 1.852

Asian Semantic Web Conference, (Springer, Berlin, 2007).

[19] X. Dong, A. Y. Halevy, and J. Madhavan, Reference reconciliation in complex information spaces. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, (ACM, New York, 2005).