Intelligent Information Retrieval in Data Mining

Ravindra Pratap Singh, Poonam Yadav

Abstract: In this paper we present the methodologies and challenges of information retrieval. We will focus on data mining, data warehousing, information retrieval, data mining ontology, intelligent information retrieval. We will also focus on fundamental concepts behind all information retrieval methods. The Data Mining Ontology is a formal description of the domain concepts and the relation between the concepts; it uses a hierarchical structure containing the concept entities and their relation. We discuss some issues and challenges facing developing the new techniques targeted to resolve some of the problems associated with intelligent information retrieval.

Keywords: intelligent Information retrieval, information extraction, data mining ontology;

1.0 DATA WAREHOUSING

Data warehousing is a process of centralized data management and retrieval. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis as described in Qi Luo [1]. A data warehouse provides the right foundation for data mining. The data warehouse makes data mining more feasible by removing many of the data redundancy and system management issues allowing users to focus on analysis.

2.0 DATA MINING

Knowledge Discovery and Data Mining are speedily developing areas of research that are at the junction of several disciplines, including statistics, databases, AI, visualization, and high-performance and parallel computing . Knowledge Discovery and Data Mining goal is to "turn data into knowledge." Basing on it, Data mining is the core part of the Knowledge Discovery in Database (KDD) process shown in Fig. 1 as described in Qi Luo [1].

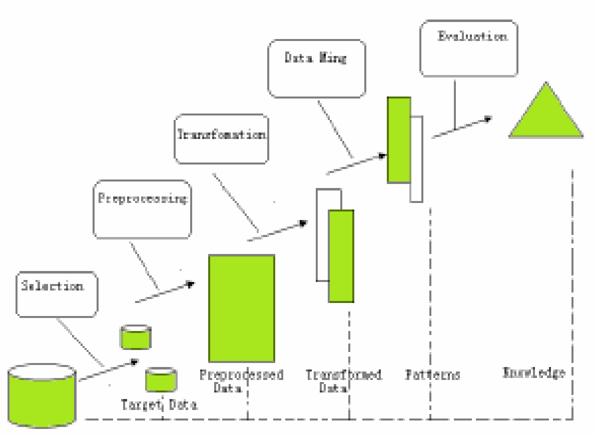

Fig. 1 Knowledge Discovery in Database (KDD) process

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

The KDD process may consist of the following steps:

1 Data selection
2 Data cleaning
3 Data transformation
4 Pattern searching (data mining)

Data mining and KDD are often used interchangeably because data mining is the key to the KDD process
as described in Qi Luo[1].

2.1 Data Mining Tasks

Different methods and techniques are needed to find different kinds of patterns. Based on the patterns tasks in data mining can be classified into Summarization, classification, clustering, association and trend Analysis as described in Qi Luo [1]

(1) Summarization. Summarization is the abstraction or generalization of data. A set of task-relevant data is summarized and abstracted. This results in a smaller set which gives a general overview of the data, usually with aggregate information.

(2) Classification. Classification derives a function or model which determines the class of an object based on its attributes.

(3) Clustering. Clustering aims to establish relatively homogeneous subgroups in the given data.

(4) Trend analysis. Time series data are records accumulated over time. Such data can be viewed as objects with an attribute time.

2.2 Data Mining Technology

Data mining adopted its techniques from many research areas including statistics, machine learning, association Rules, neural networks as described in Qi Luo [1].

(1) Association Rules. Association rule generators are a powerful data mining technique used to search through an entire data set for rules revealing the nature and frequency of relationships or associations between data entities. The resulting associations can be used to filter the information for human analysis and possibly to define a prediction model based on observed behavior.

(2) Artificial Neural Networks are recognized in the automatic learning framework as universal approximations, with massively parallel computing character and good generalization capabilities, but also as black boxes due to the difficulty to obtain insight into the relationship learned.

(3) Statistical Techniques. These include linear regression, discriminate analysis or statistical summarization.

(4) Machine learning (ML) is the center of the data mining concept, due to its capability to gain physical insight into a problem, and participates directly in data selection and model search steps.

2.3 Heuristic Classification in Data Mining

Clancy & Fayyad et al[2]. Proposed a general problem solving method called Heuristic Classification shown in Fig 2. The data mining is a goal-directed task. There are always models and patterns within the data. Constructing a successful data mining system needs to know what to look for in prior. The users must provide a well-defined problem that can be solved by the data mining algorithms.

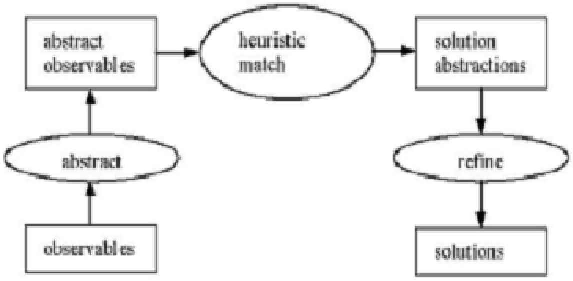

Fig. 2 The Problem solving method Heuristic classification

By the heuristic classification for data mining, the user observes the data and algorithms locate in the right side of Fig. 2. The heuristic match process matches a set of algorithms satisfying the proposed goal. As one algorithm alone normally can't solve the task, the algorithms may also need to be subdivided into sub-procures described by the refine step in Fig. The whole data mining project can be identified by the heuristic classification model with the following steps:

- Identify the mining task.
- Match the task to one solution within a solution set.
- Decompose the solution to several sub-algorithms.

3.0 The Data Mining Ontology

Ontology is a formal description of the domain concepts and the relation between the concepts; it uses a hierarchical structure containing the concept entities and their relation. Ontology would be understandable for both human beings and machine and could be reused. For constructing ontology for data mining, we need a model for modeling the data mining algorithms. We take the data mining modeling only concerns with algorithms introduced by Fayyad shown in Fig. 3.

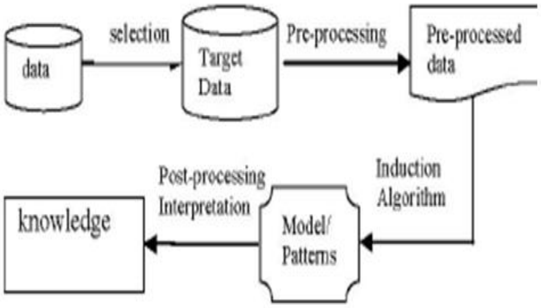

Fig. 3: The data mining process

For a selected data, firstly, pre-processing is preformed on the data, such as filling the missing values, cleaning the data, or extracting features from the data. Then one of the induction algorithms is introduced for mining the pre-processed data according to the task, for example, if the users want to find the relationship between temporal data items within a transaction database, association rule mining can be used. After that, hypothesis testing may be applied to check the important rules and the mined rules are given to the user for interpretation.

3.1 The Structure of Data Mining Ontology

Data mining process	
Data type	
Function	
Method	
Algorithm	
Constraint	Performance

Some algorithms can only be used for some specific form of data type. Thus each algorithm should be within the sub-tree of data type. For classify the structure, we arrange each algorithm in some specific function. Note some algorithms serve for multiple functions, for example, SVM can be used both for classification and regression. In this case, this algorithm will appear in each function category.

4.0 Information Retrieval (IR)

In IE, we retrieve only structured data. And IR provides all types of information access such as analysis, organization, storage, searching and retrieval of information. As according to Slaton's classic textbook**:** "Information retrieval is a field concerned with the structured, analysis, organization, storage, searching, and retrieval of information" [5].

Information retrieval is defined by Carlo Meghini et al. "Information Retrieval as the task of identifying documents in a collection on the basis of properties described to the documents by the user requesting the retrieval" [4].

Combination of IR and database will be valuable for the development of probabilistic models for integrity unstructured, semi-structured and structured data, for the design of effective distributed, heterogeneous information systems [5]. Jimmy Lin describes the information retrieval cycle as given in figure4 .

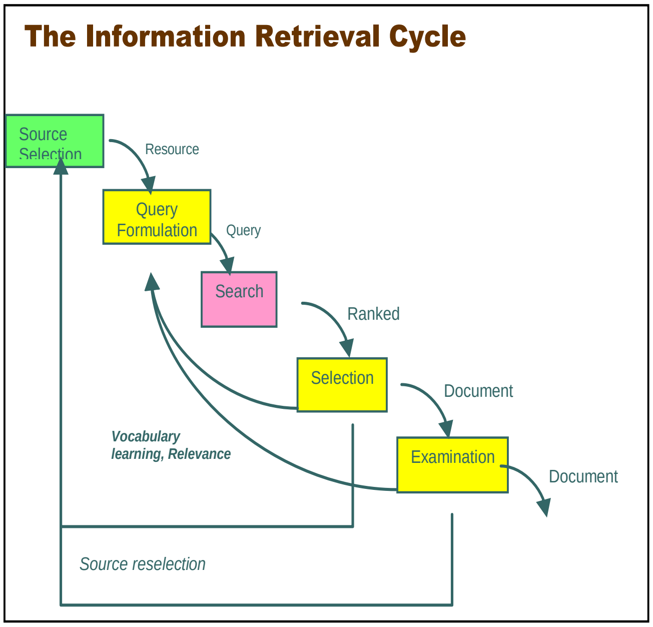

Figure 4: Information Retrieval Cycle

In figure4, source selection is first step in IR cycle. Appropriate resource is selected for all valuable data. Then query will be formed according to user needs in query formulation. Query will be processed for searching desired results, rank list will be prepared according to search results and appropriate documents are selected. If selection is not according to user needs, selection criteria goes back to source selection and query formulation. Examination process checks the documents for validation, if its validation is proved then documents will be the result otherwise output of examination goes back to source reselection and query formulation.

Various challenges in IR:

New challenges to IR community and motivated researchers to look for intelligent Information Retrieval (IR) systems that search and/or filter information automatically based on some higher level of understanding are required. What type of semantics is to be used to improve effectiveness in IR? What will be the hypothesis to improve representation of documents and by incorporating limited semantic knowledge? [7].

5.0 INTELLIGENT INFORMATION RETRIEVAL SYSTEMS

JIANG Xinhua et al described that the retrieved documents have the right terms but may not be in the desired context. It has been argued that exploiting the user's context has the potential to improve the performance of information retrieval system described in [10].

The challenge of developing intelligent information retrieval system based on context of documents:

- How to develop a methodology of assigning context to documents using assigners and validate the results through measurement of consistency between assigners?
- Web usage tools, usually implementing the process of customizing the content and the structure of web sites in order to satisfy the specific need of each and every user, without asking for it explicitly;
- Web content mining tools used for automatic classification of document contents, including their multimedia objects as well as textual information.

6.0 ONTOLOGY AND INFORMATION RETRIEVAL

JING-YAN WANG et al. described that the characteristics of an ideal information retrieval system are that searching is fleet and result is accurate described in [9]:

Ontology is the formalized description of share conceptual model to the information resource, so any conceptual object C can be defined as: C = {D, W, R} Where, D indicates such domain knowledge, W is a status set of correlative thing in applied domain; R is a set of concept relation in domain space {D, W}.

It has two effects that ontology is used in information retrieval.

(1) Automatic analysis to the domain attribute of document.
The information retrieval system looks for keywords according to a certain means in searched information document; these keywords can be used in document classification which is dependent on ontology knowledge. So information retrieval system can sort documents to some possible domains combining with primary content of information document, then filtrates off irrelevant domains, gets final result of correlative domains which the document is classed to.

(2) Intelligent formulation and visualization to user's searching demand.

It is very important that how to visualize user's knowledge demand by appropriate method in preliminary stage of information retrieval.

7.0 CONCLUSION

Data warehousing is a process of centralized data management and retrieval. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified The Data Mining Ontology is a formal description of the domain concepts and the relation between the concepts, it uses a hierarchical structure containing the concept entities and their relation. Ontology would be understandable for both human beings and machine and could be reused.

Information retrieval provides all types of information access. Concept-based visual indexing has high expressive power, which can easily communicate with user but still involves information loss in transforming visual materials into text, and requires more intensive human labor. Information Extraction is a logical step to retrieve structured data and the extracted information. The extracted information is then used in searching, browsing, querying, and mining. The retrieved documents have the right terms but may not be in the desired context. It has been argued that exploiting the user's context has the potential to improve the performance of information retrieval systems.

There are two important issues to be addressed. First one is to design and integrate an efficient ontology and the second is to establish of linkages between ontologies of different domains.

REFERENCES:

[1] Qi Luo, "Advancing Knowledge Discovery and Data Mining", 2008 IEEE DOI 10.1109/WKDD.2008.153

[2] Mao-Song Lin, Hui Zhang, and Zhang-Guo Yu, "An Ontology for Supporting Data Mining Process" Manuscript received June 30, 2006.

[3] AnHai Doan, Raghu Ramakrishna, and Shiva Kumar, "Managing Information Extraction", SIGMOD 2006, ACM 1-59593-256-9/06

[4] CARLO MEGHINI, FABRIZIO SEBASTIANI, AND UMBERTO STRACCIA, "A Model of Multimedia Information Retrieval", Journal of the ACM, Vol. 48, No. 5, September 2001, pp. 909–970.

[5] G. Salton, "Automatic Information Organization and Retrieval", McGraw-Hill, New York, 1968.

[6] Challenges In Information Retrieval and language Modeling, Report of a Workshop held at the center for intelligent Information Retrieval, University of Massachusetts Amherst, and September 2002.

[7] Jimmy Lin, "An introduction to information retrieval and question answering", 2004.

[8] Tanveer J Siddiqui "Intelligent Techniques for Effective Information Retrieval", 2006.

[9] JING-YAN WANG, ZHEN ZHU, "FRAMEWORK OF MULTI-AGENT INFORMATION RETRIEVAL SYSTEM BASED ON ONTOLOGY AND ITS APPLICATION" 978-1-4244-2096-4/08/$25.00 ©2008 IEEE

[10] JIANG Xin-Hua1, LIU Yong-Min, "A New Artificial Intelligent Information Retrieval Methods" 978-0-7695-3559-3/09 $25.00 © 2009 IEEE

[11] Jimmy Lin, "An introduction to information retrieval and question answering", 2004.

[12] Chen-Yu Lee ,Von-Wun SOO, " Ontology based Information Retrieval and Extraction" IEEE 0-7803-8932-8/05 2005 .

[13] Hao Han and Takehiro Tokuda , "A Method for Integration ofWeb Applications Based on Information Extraction" ICWE.2008. IEEE DOI 10.1109/IC

[14] Sebastian A. R´ýos_, Juan D. Vel´asquez†, Eduardo S. Vera‡§, Hiroshi Yasuda_ and Terumasa , " Improving Web Site Content Using a Concept-based Knowledge Discovery" IEEE/WIC/ACM International Conference on Web Intelligence 2006.

[15] Gilles Nachouki, "A Method for Information Extraction from the Web"0-7803-9521-2/06/$20.00)2006 IEEE.

[16] Hao Han and Takehiro Tokuda, "A Method for Integration ofWeb Applications Based on Information Extraction".

[17] Xianming Liu , Weihua Li , Shixian Li3 Zhenwen Tao , "A New Model to Describe Information Semantic" International Symposium on Information Science and Engieering,2008 IEEE

[18] Yi Xiao, Ming Xiao, Fan Zhang, "Intelligent Information Retrieval Model Based on Multi-Agents" 1-4244-1312-5/07/$25.00 © 2007 IEEE

[19] Mohan S. Kankanhalli and Yong Rui, "Application Potential of Multimedia Information Retrieval" IEEE | Vol. 96,No. 4, April 2008 0018-9219/$25.00 _2008 IEEE

[20] Pan Ying,Wang Tianjiang, Jiang Xueling, " Building Intelligent Information Retrieval System Based on Ontology" 1-4244-1135-1/07/$25.00 ©2007 IEEE.

ABOUT THE AUTHOR:

Dr. R. P. Singh received B.Tech. and Ph.D. degrees in electrical engineering from K.N.I.T; Sultanpur (U.P.) and Institute of Technology, Banaras Hindu University,Varanasi (U.P.), India, in 1988 and 1993, respectively.

Dr. Singh is a person of multidimensional personality with 23 years of experience in the field of teaching and research including 4 years as Director/ Principal. Dr. Singh is a member of various academic societies of national and international repute. He has to his credit 40 research papers and authored four books.These books are well received by the students & academicians alike. In 1994-1995 he was the recipient of "The President of India" as Director/ Principal award for research contributions. Currently he is working as Director/ Principal in Bimla Devi Educational Society's
Group of Institutions (Integrated Campus)

JB Knowledge Park, Faridabad and his email id is ravindra10765@gmail.com .

Poonam Yadav obtained B.Tech in Computer Engg.&Science from Kurukshetra University Kurukshetra, India and M.Tech in Information Technology from Guru Govind Singh Indraprastha University in 2002 and 2007 respectively. She is currently working as Assistant Professor in D.A.V College of Engg. & Technology, Kanina (Mohindergarh). Presently she is pursuing her Ph.D. research at NIMS University, Jaipur. Her research interests include Information Retrieval; Web based retrieval and Semantic Web etc. Mrs. Yadav is a life time member of Indian Society for Technical Education and her email id is poonam.ir@gmail.com .