KNOWLEDGE RETRIEVAL AND DATA MINING

Mrs. Bhavana Tiple , Sachin Dhande

Abstract — **The rapid growth and adoption of the World Wide Web has further exacerbated the user need for efficient mechanisms for information and knowledge location, selection and retrieval. Much work is required to address knowledge retrieval; for instance, users' information needs could be better interpreted, leading to accurate information retrieval. This paper is to suggest knowledge retrieval as a new research field and also proposes a knowledge retrieval model combining knowledge search with data mining technologies. In this model, data mining is integrated into the whole retrieval procedure of query optimizing, searching, results analyzing, and resources constructing. It realizes knowledge retrieval by various approaches, different levels, and multi-modes, and significantly improves knowledge retrieval level and efficiency. Furthermore, we explore related knowledge retrieval methods and algorithms including association analysis-based concept retrieval and inductive learning-based classification retrieval.**

Key Words — **Knowledge retrieval; Data mining; Retrieval model; Retrieval algorithm.**

I. INTRODUCTION

The Knowledge Environment represented by Web2.0 and semantic web, promotes internet into a globalized sharing knowledge networks. It implies massive static/dynamic information and knowledge resources, covering both the outer explicit objective knowledge and the inner implicit subjective knowledge. Knowledge-driven network environment, on the one hand, provides a wealth of knowledge resources for retrieval; on the other hand, how to search and discover the valuable knowledge deeply from plentiful resources. Traditional information retrieval technologies they are based on the syntax-level information organization which cannot fully express the semantic contents and relations about resources. Its retrieval mechanism commonly uses the retrieval pattern of word matching. Most search systems return a great deal of irrelevant information or lose important information [5]. Therefore, it is a new challenge that knowledge retrieval is accomplished by the integrated application with semantic retrieval, data mining, ontology.

II. KNOWLEDGE RETRIVAL SYSTEM

A. Comparison of Retrieval Systems

Traditional information retrieval systems lack the management at the knowledge level. A user must read and analyze the relevant documents in order to extract the useful knowledge. Knowledge Retrieval Systems (KRS) is the next generation retrieval systems for supporting knowledge discovery, organization, storage, and retrieval. Such systems will be used by advanced and expert users to tackle the challenging problem of knowledge seeking [3].

Knowledge retrieval (KR) focuses on the knowledge level. We need to examine how to extract, represent, and use the knowledge in data and information. Knowledge retrieval systems provide knowledge to users in a structured way. They are different from data retrieval systems and information retrieval systems in inference models, retrieval methods, result organization, etc.

The core of data retrieval and information retrieval are retrieval subsystems. Data retrieval gets results through Boolean match. Information retrieval uses partial match and best match. Knowledge retrieval is also based on partial match and best match. Considering inference perspective, data retrieval uses deductive inference, and information retrieval uses inductive inference. Considering the limitations from the assumptions of different logics, traditional logic systems cannot make efficient reasoning in a reasonable time. Associative reasoning, analogical reasoning and the idea of unifying reasoning and search may be effective methods of reasoning at the web scale.

From retrieval model perspective, knowledge retrieval systems focus on the semantics and knowledge organization. Data retrieval and information retrieval organize the data and documents by indexing, while knowledge retrieval organizes knowledge by connections among knowledge units and the knowledge structures [3].

B. A Typical KR System

Knowledge represented in a structured way is consistent with human thoughts and is easily understandable. Sometimes, users do not know exactly what they want or are lack of contextual awareness. If knowledge can be provided visually in a structured way, it will be very useful for users to explore and refine the query. Fig1 shows a conceptual framework of a typical knowledge retrieval system. The main process can be described as follows:

1) *Knowledge Discovery:* Discovering knowledge from sources by data mining, machine learning, knowledge acquisition and other methods.

2) *Query Formulation:* Formulating queries from user needs by user inputs. The inputs can be in natural languages and artificial languages.

3) Knowledge Selection: Selecting the range of possible related knowledge based on user query and knowledge discovered from data/information sources.

4) Knowledge Structure Construction: Reasoning according to different views of knowledge, domain knowledge, user background, etc. in order to form knowledge structures. Domain knowledge can be provided by expert systems. User background and preference can be provided by user logs.

5) Exploration and Search: Exploring the knowledge structure to get general awareness and refine the search. Through understanding the relevant knowledge structures, users can search into details.

6) *Knowledge Structure Reorganization:* Reorganizing knowledge structures if users need to explore other views of selected knowledge.

7) *Query Reformulation:* Reformulating the query if the constructed structures cannot satisfy user needs. One of the key features of knowledge retrieval systems is that knowledge are visualized in a structured way so that users could get contextual awareness of related knowledge and make further retrieval [6].

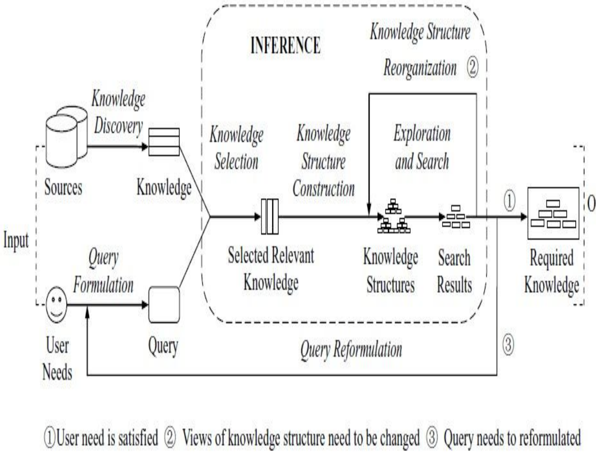

Fig 1. A Typical KR System

III. Knowledge Retrieval And Data Mining

Knowledge retrieval is developed gradually on the basis of information retrieval, and is to provide the information and knowledge Ontology-based knowledge organization can well express the contents of information elements and the various semantic relations between them and support semantic reasoning and retrieval [4]. Data mining, ontology and semantic technology provide an effective approach for knowledge retrieval at semantic-level [1]. Data mining is the process of acquiring hidden, unknown, and potentially useful information and knowledge from a large number of data [1],[7]. Data mining has broad future applications and has attracted much attention in the academic community and business circle. The research objectives of knowledge retrieval and data mining are the same in essence [1].They complement each other and can achieve the in-depth knowledge acquisition from information resources.

A. *Knowledge Retrieval Model*

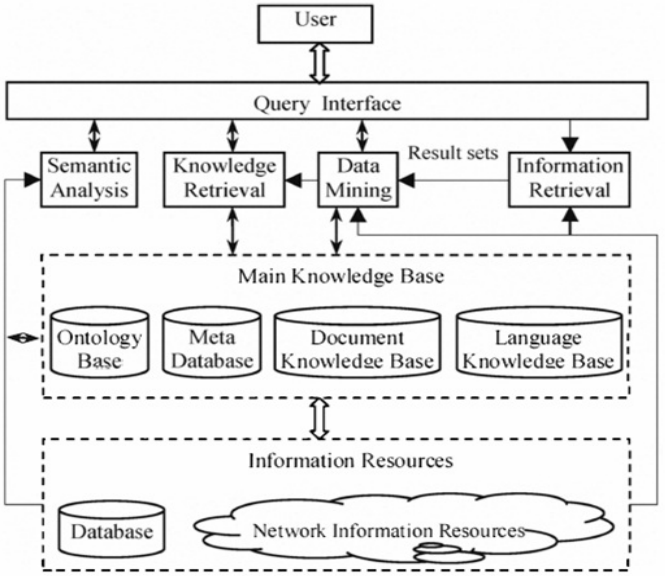

Fig 2: Knowledge Retrieval Model

The main components of the proposed knowledge retrieval model are showed in Fig 2. They are semantic analysis, knowledge retrieval, data mining, information retrieval and knowledge base.

1)Semantic Analysis: The main function of this part is to analyse the semantics of user questions, search results and database by using ontological knowledge. Then it builds the semantic query expression composed of concepts and its associations. At the same time, the knowledge base is constructed and improved after the semantic analysis and annotation of search results and information elements. Rich semantic information is hidden in retrieval context [1],[8].

2) Knowledge Retrieval: The retrieval mechanism realizes the retrieval and mining from document knowledge base or information resources. The main retrieval methods include concept retrieval based on association analysis, classification retrieval based on inductive learning, ontology association retrieval and semantic reasoning retrieval, while the information retrieval component acts as an auxiliary mechanism for knowledge retrieval.

3*) Data Mining:* Data mining is applied into the entire process of query optimizing, searching, result analysing, and knowledge resources constructing. Its main functions include: Do mining processing to the results of information retrieval. The processing includes classification analysis, inductive learning, association analysis, cluster analysis, automatic summarization and information extraction. Then, the mining results can be presented to the user or be used to optimize the user queries and expand knowledge base. Integrate data mining into the searching process. Mine the knowledge directly from the Meta database, document knowledge base, database and network information resources.

4) *Knowledge Bases:* Ontology Base covers all the high-level abstract concepts of knowledge resources, the relationships between these concepts and the description of the core knowledge patterns, which is also known as the concept base.

Meta Database contains all the metadata descriptions of information resources. It is the foundation for knowledge retrieval and mining. Documents Knowledge Base including the semantic description of information elements and their relationship, it is the main resources for knowledge retrieval. Language Knowledge Base is used for the storage of linguistic knowledge required by system, mainly including dictionaries, grammar and semantic knowledge, to support semantic analysis and process.

B. *Knowledge Retrieval Strategy and its Implementation Process*

Knowledge retrieval model provides various retrieval strategies and approaches.

1) Information Resource Retrieval. The retrieval results are analysed and mined, and provided to users or used to optimize user queries;

2) Retrieval on Concept Knowledge Base: It is combined with mining technology to obtain the concept knowledge.

3) Retrieval on Document Knowledge Base: The heuristic knowledge of mining is applied to guide high-efficient classification search;

4) Directly Apply Data Mining Technologies: Apply it *to* discover knowledge from the search results, ontology base, document knowledge base, and database and network information resources, to support and optimize the knowledge retrieval processing. Specifically, the whole knowledge retrieval process includes semantic matching, analysis, retrieval, mining and optimization. After inputting user query, the system first analyses its semantics and describes the user search concepts and the semantic associations which will be organized into a semantic network. Then, the query will be submitted to the information retrieval or knowledge retrieval mechanism to search the database or knowledge base. At last, the knowledge, rules, patterns, summary reports, and satisfying documents are presented to the user. At the same time, related knowledge is added to knowledge base.

C. *ONTOLOGY-BASED RETRIEVAL FRAMEWORK*

The ontology-based knowledge retrieval framework consists of four models: a user’s mental model and querying model, a computer model, and an ontology model. A user’s mental model is her (his) background knowledge system. A querying model is a user’s translation of an information need generated from her (his) mental model. The computer model constructs ontology for the user. The constructed ontology is the ontology model aiming to simulate the user’s mental model in IR [2].

1) Framework of the Knowledge Retrieval Model: The knowledge retrieval model proposed in this paper aims to acquire and analyse a Web user's background knowledge so that his (her) information need can be better captured and satisfied, two knowledge resources are used in Bthe model, first is World Knowledge Base, which provides a frame of world knowledge for a user to identify the positive and negative knowledge corresponding to an information need. The world knowledgebase also defines the backbone of a user's personalized subject ontology; other knowledge resource is Local Instance Repository, which provides a resource to discover a user's real information need. The framework of the knowledge retrieval model is presented in Fig. 3. The model takes a query from a user, extracts a set of potentially relevant subjects from the world knowledge base, and displays the subjects to the user, The user identifies the related knowledge including positive and negative subjects from the present subjects. Finally, based on the user identified knowledge, the model constructs a subject ontology, as the partial ontology illustrated in Fig.3. Once a user's subject ontology is constructed, the knowledge for user information needs can then be mined from the user's LIR and the constructed ontology.

The model produces a set of subjects related to a user's interests and helping to interpret the user's information need. Our proposed knowledge retrieval model uses ontology’s to specify a user's background knowledge and to capture a user's information need. This model attempts to enhance existing IR techniques by solving problems on the knowledge level, and to fill the related research gap in the IR development [3].

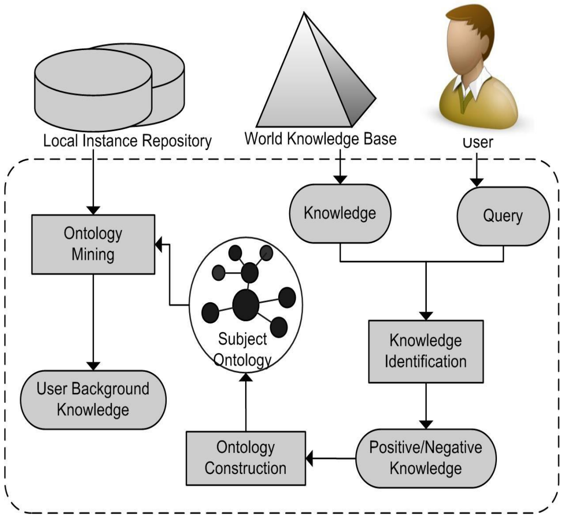

Fig 3: Framework of the Knowledge Retrieval Model[3]

IV. KNOWLEDGE RETRIEVAL METHODS AND ALGORITHMS

This topic will focus on the integration of classification analysis, machine learning, association analysis into knowledge retrieval, and research the concept retrieval based on association analysis and the classification retrieval methods based on inductive learning. In the process of concept retrieval, a multi-level association rule mining algorithm is introduced to obtain deep concepts and semantic association knowledge. In the process of classification retrieval, the classification analysis on the results of information retrieval can produce feature description to optimize retrieval processing.

A. *Concept retrieval algorithms based on association analysis*

Concept retrieval is achieved by concept processing and semantic association analysing. there are associations between concepts, such as synonym, near-synonym, upper and lower, etc. Usually, kinds of concepts and their associations are acquired by experts or by using data mining technologies. Concept base is often organized as a hierarchical semantic network. It can provide users with high-quality concepts and their semantic association knowledge. However, it is difficult to meet various requirements of users of wide range, and probably fails in inquiry. To this end, we treat the user queries as constraints for the further multi-level association analysis of the meta database to discover related concepts and associations. If the search result is satisfying for the user, it will be credited into the learning set of concept base to realize the real-time or periodic dynamic learning. The mining of meta database as a supplementary method of concept retrieval, not only can meet the broad requirements of more users, but also may obtain surprises with the application of multi-level mining. Although the mining of meta database is slower than the searching of concept base, it does not need lengthy mining for large databases, thus, it can improve the retrieval efficiency.

The following is the step of multilevel association analysis mining.

1*) Construct the concept hierarchy tree:*. In fact, the meta database has a good hierarchy, just needing some simple processing.

2) *Normalize and generalize the concepts:* Concept normalization is to convert the concepts of different levels in a certain attribute into those at the same level. Concepts generalization is to convert the low-level concepts into high level concepts. The standardization and generalization of concept can be implemented with the relations of the super and the sub classes. For example, use the relations of superClassOf, subClassOf and same As for reasoning to normalize and generalize the concepts of different levels.

3*) Find the frequent item set*: The so-called frequent item set is that the frequency of the items in the set is higher than or at least equal to the predefined minimum support. The "support-confidence" framework can be basically adopted in multi-level association rule mining. Generally speaking, top down strategy can be applied. It starts from the highest concept level down to a certain lower level to do accumulated counting of the frequent item sets at each level until it can't find one more. For each level, the existing mining algorithms can be used to find frequent item sets.

4) *Generate the strong association rules with frequent item sets*: The strong rule is the rule which can meet both minimum support threshold and minimum confidence threshold.

5) *Remove redundant and useless rules*: In the process of multi-level association rule mining, some rules may be redundant, and should be deleted.

B. *Classification retrieval algorithms based on inductive learning*

Classification retrieval method is rested on the query concept to search the class hierarchy of document knowledge base or database directly. In order to improve the search efficiency, classification analysis mining methods are utilized in the process of knowledge retrieval. First, information retrieval is performed based on the query concepts. Then, inductive learning methods are employed for the classification analysis on search result set, to get a satisfying document set and its classification feature descriptions.

The classification retrieval algorithm based on inductive learning is as follow:

1) *Information search*: Search information resources and get a retrieval result set.

2) *Inductive learning:*. Do classification analysis on the retrieval result set to obtain classification feature descriptions:

First divide the retrieval result set into two sets: positive instance set and negative instance set then Convert positive instance set and negative instance set and represent them as training set after that apply inductive learning algorithm to training set and generate the classification feature descriptions of positive instance set.

3) *Expand and improve query*. Exact core concepts from the classification feature description are added to query expression.

4) *Do classification search: First* find document knowledge base or database level by level from up to down by the heuristic knowledge, until find the lowest subclass which matches with or contain query concepts. According to the query concepts, find all the instances of this subclass. If success, turn 5); or turn to 1) .

5) *Sort search result set*: Select and return the search results.

V. CONCLUSION AND FUTURE SCOPE

The main contribution for this paper is to suggest knowledge retrieval as a new research field. Integrating data mining technologies into the entire process of knowledge retrieval can provide kinds of retrieval strategies and methods and support dynamic learning, mining and adaptability. It can realize deep knowledge discovery and retrieval from information resources so as to raise the level of knowledge retrieval and prompt its efficiency. And for the same reason, it will also promote the rapid development of the research and application of data mining. An ontology-based knowledge IR framework is proposed aiming to discover a user's background knowledge to improve IR performance.

Pure ontology-based model performs poorly when ontology information is not available or incomplete. Dynamic characteristics of the marketing make KDDM difficult to deal with unstructured data. In addition, theoretical and applied data mining research on network marketing still at an early exploratory stage, the dynamic characteristics add various difficulties to the study.

REFERENCES

[1] YanHao, Yu-feng Zhang, " Research on Knowledge Retrieval by Leveraging Data Mining Techniques", 2010 International Conference on Future Information Technology and Management Engineering.

[2] Xiaohui Tao, Yuefeng Li, Ning Zhong_, Richi Nayak," An Ontology-based Framework for Knowledge Retrieval", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[3] Xiaohui Tao_, Yuefeng Li, and Richi Nayak, " A Knowledge Retrieval Model Using Ontology Mining and User Profiling"

[4] K.Z.Gao,Z.Q.Bao,X.Q.Li, "Study on Two-layer Knowledge Retrieval Technology in Conceptual Design," Grey Systems and Intelligent Services, 2007,GSIS 2007,pp.1523-1527,2007.

[5] L.M.Shao,S.G.Zhang,X.S.Suo, "Research on Ontology Knowledge Retrieval to the Expert Consultation of Greenhouse," Communication Technology, 2006, ICCT '06,pp.1-4,2006.

[6] Yiyu Yao, Yi Zeng, Ning Zhong, Xiangji Huang. "Knowledge Retrieval (KR)", In: Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Silicon Valley, USA, November 2-5, 2007, 729-735.

AUTHOR'S PROFILE

Prof. Bhavana Tiple

Prof. Bhavana Tiple received her the M.E.degree in Computer Engg from Pune University 2005.Currently she is working as Associate Professor in MIT Engg College, Pune, India. Presently she also received grant from BCUD Pune for her research work in Domain Ontology and knowledge retrieval. Her research interests are in the areas of Data mining and Domain Ontology,knowledge retrieval,and Mulitmedia Information retreival.

Sachin Dhande

Sachin Dhande received his B.E.degree in Computer Science & Engg from Sant Gadge Baba Amravati University 2011.Currently he is Pursuing M.E. in Computer Engg from MIT Engg College, Pune University, India. His research interests are in the areas of Data mining and Mulitmedia Information retreival.