

Intelligent Query Answering

Zbigniew W. Ras

University of North Carolina, Charlotte, USA

Agnieszka Dardzinska

Bialystok Technical University, Poland

INTRODUCTION

One way to make query answering system (QAS) intelligent is to assume a hierarchical structure of its attributes. Such systems have been investigated by Cuppens & Demolombe (1988), Gal & Minker (1988), and Gaasterland et al. (1992), and they are called cooperative. Any attribute value listed in a query, submitted to cooperative QAS, is seen as a node of the tree representing that attribute. If QAS retrieves an empty set of objects, which match query q in a target information system S , then any attribute value listed in q can be generalized and the same the number of objects that possibly can match q in S can increase. In cooperative systems, these generalizations are usually controlled by users.

Another way to make QAS intelligent is to use knowledge discovery methods to increase the number of queries which QAS can answer: knowledge discovery module of QAS extracts rules from a local system S and requests their extraction from remote sites (if system is distributed). These rules are used to construct new attributes and/or impute null or hidden values of attributes in S . By enlarging the set of attributes from which queries are built and by making information systems less incomplete, we not only increase the number of queries which QAS can handle but also increase the number of retrieved objects.

So, QAS based on knowledge discovery has two classical scenarios that need to be considered:

- In a standalone and incomplete system, association rules are extracted from that system and used to predict what values should replace null values before queries are answered.
- When system is distributed with autonomous sites and user needs to retrieve objects, from one of these sites (called client), satisfying query q based on attributes which are not local for that site, we search for definitions of these non-local attributes at remote sites and use them to approximate q (Ras, 2002; Ras & Joshi, 1997; Ras & Dardzinska, 2004).

The goal of this article is to provide foundations and basic results for knowledge discovery-based QAS.

BACKGROUND

Modern query answering systems area of research is related to enhancements of query answering systems into intelligent systems. The emphasis is on problems in users posing queries and systems producing answers. This becomes more and more relevant as the amount of information available from local or distributed information sources increases. We need systems not only easy to use but also intelligent in answering the users' needs. A query answering system often replaces human with expertise in the domain of interest, thus it is important, from the user's point of view, to compare the system and the human expert as alternative means for accessing information.

Knowledge systems are defined as information systems coupled with a knowledge base simplified in Ras (2002), Ras and Joshi (1997), and Ras and Dardzinska (1997) to a set of rules treated as definitions of attribute values. If information system is distributed with autonomous sites, these rules can be extracted either from the information system, which is seen as local (query was submitted to that system), or from remote sites. Domains of attributes in the local information system S and the set of decision values used in rules from the knowledge base associated with S form the initial alphabet for the local query answering system. When the knowledge base associated with S is updated (new rules are added or some deleted), the alphabet for the local query answering system is automatically changed. In this paper we assume that knowledge bases for all sites are initially empty. Collaborative information system (Ras, 2002) learns rules describing values of incomplete attributes and attributes classified as foreign for its site called a client. These rules can be extracted at any site but their condition part should use, if possible, only terms that can be processed by the query-answering system associated with the client. When

the time progresses more and more rules can be added to the local knowledge base, which means that some attribute values (decision parts of rules) foreign for the client are also added to its local alphabet. The choice of which site should be contacted first, in search for definitions of foreign attribute values, is mainly based on the number of attribute values common for the client and server sites. The solution to this problem is given in Ras (2002).

MAIN THRUST

The technology dimension will be explored to help clarify the meaning of intelligent query answering based on knowledge discovery and chase.

Intelligent Query Answering for Standalone Information System

QAS for an information system is concerned with identifying all objects in the system satisfying a given description. For example, an information system might contain information about students in a class and classify them using four attributes of “hair color,” “eye color,” “gender,” and “size.” A simple query might be to find all students with brown hair and blue eyes. When an information system is incomplete, students having brown hair and unknown eye color can be handled by either including or excluding them from the answer to the query. In the first case we talk about optimistic approach to query evaluation while in the second case we talk about pessimistic approach. Another option to handle such a query would be to discover rules for eye color in terms of the attributes hair color, gender, and size. These rules could then be applied to students with unknown eye color to generate values that could be used in answering the query. Consider that in our example one of the generated rules said:

$$(\text{hair, brown}) \wedge (\text{size, medium}) \rightarrow (\text{eye, brown}).$$

Thus, if one of the students having brown hair and medium size has no value for eye color, then the query answering system should not include this student in the list of students with brown hair and blue eyes. Attributes hair color and size are classification attributes and eye color is the decision attribute.

We are also interested in how to use this strategy to build intelligent QAS for incomplete information systems. If a query is submitted to information system *S*, the first step of QAS is to make *S* as complete as possible. The approach proposed in Dardzinska & Ras

(2003b) is to use not only functional dependencies to chase *S* (Atzeni & DeAntonellis, 1992) but also use rules discovered from a complete subsystem of *S* to do the chasing.

In the first step, intelligent QAS identifies all incomplete attributes used in a query. An attribute is incomplete in *S* if there is an object in *S* with incomplete information on this attribute. The values of all incomplete attributes are treated as concepts to be learned (in a form of rules) from *S*.

Incomplete information in *S* is replaced by new data provided by Chase algorithm based on these rules. When the process of removing incomplete values in the local information system is completed, QAS finds the answer to query in a usual way.

Intelligent Query Answering for Distributed Autonomous Information Systems

Semantic inconsistencies are due to different interpretations of attributes and their values among sites (for instance one site can interpret the concept “young” differently than other sites). Different interpretations are also due to the way each site is handling null values. Null value replacement by values suggested either by statistical or knowledge discovery methods is quite common before a user query is processed by QAS.

Ontology (Guarino, 1998; Sowa, 1999, 2000; Van Heijst et al., 1997) is a set of terms of a particular information domain and the relationships among them. Currently, there is a great deal of interest in the development of ontologies to facilitate knowledge sharing among information systems.

Ontologies and inter-ontology relationships between them are created by experts in the corresponding domain, but they can also represent a particular point of view of the global information system by describing customized domains. To allow intelligent query processing, it is often assumed that an information system is coupled with some ontology. Inter-ontology relationships can be seen as semantical bridges between ontologies built for each of the autonomous information systems so they can collaborate and understand each other.

In Ras and Dardzinska (2004), the notion of optimal rough semantics and the method of its construction have been proposed. Rough semantics can be used to model semantic inconsistencies among sites due to different interpretations of incomplete values of attributes. Distributed chase (Ras & Dardzinska, 2004) is a chase-type algorithm, driven by a client site of a distributed information system (DIS), which is similar to chase algorithms based on knowledge discovery and presented in

3 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/intelligent-query-answering/10675

Related Content

Customized Recommendation Mechanism Based on Web Data Mining and Case-Based Reasoning

Jin Sung Kim (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 2659-2672).

www.irma-international.org/chapter/customized-recommendation-mechanism-based-web/7790

Intelligent Data Analysis

Xiaohui Liu (2005). *Encyclopedia of Data Warehousing and Mining* (pp. 634-638).

www.irma-international.org/chapter/intelligent-data-analysis/10674

Ontology-Based Data Warehousing and Mining Approaches in Petroleum Industries

Shastri L. Nimmagadda and Heinz Dreher (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 1901-1925).

www.irma-international.org/chapter/ontology-based-data-warehousing-mining/7740

Data Mining Techniques and Medical Decision Making for Urological Dysfunction

N. Sriraam, V. Natasha and H. Kaur (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 2506-2516).

www.irma-international.org/chapter/data-mining-techniques-medical-decision/7779

Re-Sampling Based Data Mining Using Rough Set Theory

Benjamin Griffiths and Malcolm J. Beynon (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 3005-3026).

www.irma-international.org/chapter/sampling-based-data-mining-using/7818