Expert Systems with Applications 40 (2013) 2109–2116

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data summarization ontology-based query processing

Hai Wang [a,1], Shouhong Wang [b,*]

[a] Sobey School of Business, Saint Mary's University, 903 Robie Street, Halifax, NS, Canada B3H 2W3
[b] Charlton College of Business, University of Massachusetts Dartmouth, 285 Old Westport Road, Dartmouth, MA 02747-2300, USA

ARTICLE INFO

Keywords:
Data summarization
Ontology
Query
Ontology-based query
Data mining

ABSTRACT

Data summarization has recently received considerable attention in the knowledge systems community. This paper discusses the design of data summarization query system. Based on an initial analysis of requirement representations in data summarization, the study develops a generic organization of ontology for data summarization query system. Furthermore, this paper proposes a framework of ontology-based query language of data summarization based on the proposed ontology structure. A prototype project of data summarization ontology-based Query by Examples (QBE) for summarizing the data incompleteness demonstrates the effectiveness of the proposed framework.

1. Introduction

Data summarization has become an important technology as the amount of data from various sources has rapidly increased (Bayam, Liebowitz, & Agresti, 2005; Garrido & Requena, 2012; Song, Choi, Park, & Ding, 2011; Yager & Petry, 2006). Many database applications require quick summative answers from the large databases. In most cases, exact answers may not be necessary but prompt feedbacks can be helpful. The central issue of data summarization is to construct a concise and sufficiently accurate approximate representation of data summary or synopsis promptly. For example, to evaluate the quality of a very large database, one must know how many data items are missing from the database, what attributes have the most missing values, and what the correlations between those missing data exist, etc. It is not unusual that the very large database runs ordinary SQL queries for days or even longer to provide accurate summarization results. In contrast, a good approach to such data summarization practices is to have a special data summarization system that can provide rapid results based on the user's realistic needs.

Recently, while a variety of data search engine technologies have been applied widely in the information technology field, few well-established data summarization technologies beyond standard statistical techniques have been available on the market. Technically, the use and the development of data summarization techniques are much more complicated than that of keywords-based search engines. In general, the current data engineering technology is underdeveloped in the aspects of semantic models, application context, and the way of presentation results (Bontcheva et al., 2006; Vesin, Ivanović, Klašnja-Milićević, & Budimac, 2012).

In this study, data summarization refers to the data processing techniques that provide derived data, metadata of the database, patterns of the data, or properties of the data, which can be approximate answers to queries based on the user's specific requirements (Wang & Sevcik, 2008; Wang & Wang, 2010). Given the complexity of data summarization applications and techniques, a special system is needed to provide the relevant, meaningful, and prompt data summarization results for the user and such a computing system is called data summarization query system.

There have been many computing techniques for data summarization (Wang, 2004), and each of data summarization techniques has its unique application context and specific objective. However, challenges remain if there is a lack of unified organizational framework that integrates the diversified data summarization techniques for various applications to provide comprehensible services to all types of users.

Presumably, the semantic technology can be beneficial for sharing common understanding of the context and techniques of data summarization among the various users (Kroeker, 2010). However, little research into semantic data summarization query system has been reported. In this paper, the context of this study is discussed through a literature review. A generic organization of ontology for the data summarization domain is outlined. Ontology-based query constructs is then proposed. A case study of ontology-based query processing is presented. The case study demonstrates a prototype of ontology-based data summarization query system and an example of the effective use of ontology for data summarization query. Finally, the contributions of this study are summarized.

* Corresponding author. Tel.: +1 508 999 8579; fax: +1 508 999 8646.
E-mail addresses: hwang@smu.ca (H. Wang), swang@umassd.edu (S. Wang).
[1] Tel.: +1 902 496 8231; fax: +1 902 496 8101.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.012

2. Related work

2.1. Ontology

Ontology is a science that studies explicit formal specifications of the resources and relationships among them in the domain (Gruber, 1993). An ontology is a specification of a conceptualization (Gruber, 1995), and intended for knowledge sharing among applications (Welty, 2003). Technically, an ontology is a graph/network that consists of (1) a set of concept (vertices in a graph); (2) a set of relationships connecting the concepts (directed edges in a graph; and (3) a set of instances assigned to particular concepts (data records assigned to the concepts) (Goldstone & Rogosky, 2002; Grobelnik & Mladenic, 2006; OWL, 2012).

There have been a variety of application areas in the ontology technology field (Davies, Studer, & Warren, 2006; Sanchez, Montserrat Batet, Isern, & Valls, 2012; Wang & Wang, 2008; Wu, Lin, Jiang, & Wu, 2011). This study uses ontology as a design instrument for developing a data summarization query system. Specifically, ontology in this study is a schema that organizes resources into a semantic network for data summarization applications. A crucial issue in the development and the use of ontology for information system enabled knowledge sharing is the visualization of ontology through the support of human-computer interface (Euzenat & Shvaiko, 2007). The visualization of ontology allows people, who have unique experiences and thus have different semantic networks, to share common concepts. The standard ontology modeling language (OWL, 2012) is an XML style tool and supports majorly information processing automation, but does not facilitate the user-computer interface development. There have been several ontology visualization methods (Katifori, Halatsis, Lepouras, Vassilakis, & Giannopoulou, 2007). However, the visualization of ontology is subject domain dependent. For instance, the visualization method for an ontology of biochemistry can hardly be applied to the computer information systems area. Although there have been countless methodologies of ontology development in various domains, few ontology visualization methods that are potential useful for the data summarization area can be found in the literature.

2.2. Ontology-based query

Query tools are characterized by structured query language (SQL), the standard query language, or data retrieval user interface, for relational database systems. The traditional SQL supports information retrieval without taking account of application context. Recently, research (Bobillo, Delgado, & Gómez-Romero, 2008; Dong, Yang, & Su, 2011; Dragoni, Pereira, & Tettamanzi, 2012) has indicated that effective information retrieval must combine search technologies and application context into a single framework in order to provide the most appropriate answer for user's information needs. A contextual query framework relies on various resources of the application background and environment for optimal retrieval accuracy and efficiency. Ontologies provide rich semantic concept and can be a base for contextual query frameworks (Liu & Yu, 2004). An ontology-based query construct defines query operations that operate the resources represented by the domain ontology to retrieve the interested information (Andreasen & Bulskov, 2009; Knappe, Bulskov, & Andreasen, 2007). An ontology-based query system is powered with semantic information formalized in the ontology so that it is capable to integrate all computational resources in information retrieval (Savvas & Bassiliades, 2009).

The ultimate objective of information retrieval from the database for data summarization is the formulation of data summaries through the use of a variety of techniques in a variety of application context. A comprehensive query system for all purposes of data summarization is yet to be established. As discussed later in this paper, data summarization involves other resources in addition to the database, and the concept of data summarization must be represented by the ontology. Next, we discuss the key components of ontology for data summarization, and develop a framework of ontology-based queries that assist the data user to obtain specific data summaries through data summarization processes.

3. Ontology for data summarization

The ontology developed in this study plays dual roles: (1) the semantic network of data summarization applications for integrating the resources of data summarization; (2) the underlying organization of the interactive user-computer interface of a data summarization query system.

3.1. Resource categories of data summarization

Ontologies are user-driven rather than innate (Wang & Wang, 2008). An analysis of ontology for data summarization is the identification and formalization of fundamental resources of data summarization and their relationships. The initial analysis is to propose generic resource categories for the domain of data summarization based on the premise that taxonomy of formalized generic resource categories can help people to better understand and share the ontologies (Welty, 2003). The following generic resource categories for data summarization query system are extracted from the moderate literature on data summarization (see Wang, 2004).

1. *Data summarization task*: A task of data summarization is to discover meaningful metadata of the database, derived data, patterns of the data, or properties of the data for the user. A data summarization task can be formally described as a hierarchical structure of its sub-tasks of data summarization.
2. *Data*: Data is the key resource in data summarization. In the relational database environment, tables, data cubes and any SQL results are the data resource.
3. *Procedure*: A procedure is a set of formalized process sequences and instructions that can be used for the user to accomplish the data summarization task. A procedure can include formatting data, data summarization route, and result presentation.
4. *Instrument*: An elementary model that can be used for the user to obtain the wanted data summarization results is called an instrument. An instrument could a statistical tool, an artificial intelligence model (e.g., neural networks), or a specific algorithm for data summarization.
5. *User profile*: It is important for a data summarization query system to keep the profile of each user so that the data summarization query system can be optimized for the individual. A user profile includes the characteristics of the user, the user's special preferences, and records of the use of the data summarization query system.
6. *Reference*: A free-format document describing the resources discussed above and terminologies is called a reference. References provide detailed information of resources used in the data summarization for the user. For example, a reference statement of a data summarization task might advise the user that the data summarization query result is a re-used result with dated data set unless the user wants to re-run the summarization query with updated data set
7. *Relationships between resources*: The generic semantic relationships between the six categories of resources are quite simple in this application domain, although the semantic network of the instances of the resources can be complex.