Available online at www.sciencedirect.com
SCIENCE DIRECT
Information and Software Technology 48 (2006) 901–914
INFORMATION AND SOFTWARE TECHNOLOGY
www.elsevier.com/locate/infsof

Extending the UML concepts to transform natural language queries with fuzzy semantics into SQL ☆

Frank S.C. Tseng [a,*], Chun-Ling Chen [b]

[a] Department of Information Management, National Kaohsiung First University of Science and Technology, 1 University Road, YenChao, Kaohsiung 824, Taiwan, ROC

[b] Department of Computer Science and Information Engineering, National Chiao Tung University, HsinChu 300, Taiwan, ROC

Received 16 May 2004; received in revised form 20 November 2005; accepted 6 December 2005
Available online 30 January 2006

Abstract

Database applications tend toward getting more versatile and broader to comply with the expansion of various organizations. However, naïve users usually suffer from accessing data arbitrarily by using formal query languages. Therefore, we believe that accessing databases using natural language constructs will become a popular interface in the future. The concept of object-oriented modeling makes the real world to be well represented or expressed in some kinds of logical form. Since the class diagram in UML is used to model the static relationships of databases, in this paper, we intend to study how to extend the UML class diagram representations to capture natural language queries with fuzzy semantics. By referring to the conceptual schema throughout the class diagram representation, we propose a methodology to map natural language constructs into the corresponding class diagram and employ Structured Object Model (SOM) methodology to transform the natural language queries into SQL statements for query executions. Moreover, our approach can handle queries containing vague terms specified in fuzzy modifiers, like 'good' or 'bad'. By our approach, users obtain not only the query answers but also the corresponding degree of vagueness, which can be regarded as the same way we are thinking.

Keywords: Natural Language Query; UML; Class Diagram; Object-Oriented Modeling; Fuzzy Set Theory

1. Introduction

In today's highly challenging environments, knowledge is becoming an important organizational asset that enables sustainable competitive advantage. The concept of Knowledge Management (KM) [1] is that organizational users may make use of knowledge to be more effective and productive in their work. Indeed, knowledge is of limited organizational value if it is not properly shared [2]. That makes many organizations are developing information systems specifically to facilitate the sharing and integration of knowledge.

To share information and knowledge, it is inevitable to pose queries on database systems, as they are ubiquitous and popular in storing enterprise data for various applications. Although the rapid evolution of Internet enables people to share information everywhere at anytime by retrieving data from database systems, however, naïve users still suffer from issuing formal query statements arbitrarily, which may frustrate users and limit the information sharing process.

Since most of the human knowledge is recorded in linguistic form, systems that could understand natural languages could help to access various kinds of information. Therefore, one of the most natural ways to issue queries on databases is using natural languages. In particular, as we have shown in [23], many retrieval tasks posed by complex formal queries can be expressed by using only simple natural language statements.

E-R diagrams was introduced by Chen [7] in 1976 and had its widely usage to model the entities within a relational database schema. For mapping between natural language and database schema, Chen [6] has addressed 11 rules for translations between natural language constructs and E-R diagrams, and our prior research [23] proposed a methodology to map natural language constructs into relational algebra through E-R representations.

Isoda [11] has addressed that Object-Oriented (OO) analysis approach allows us to intuitively and naturally model the real

☆ This research was partially supported by National Science Council, Taiwan, ROC, under Contract No. NSC 94-2416-H-327-009.

* Corresponding author. Tel.: +886 7 6011000; fax: +886 7 7659541.
E-mail addresses: imfrank@ccms.nkfust.edu.tw (F.S.C. Tseng), gis93804@cis.nctu.edu.tw (C.-L. Chen).

0950-5849/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2005.12.002

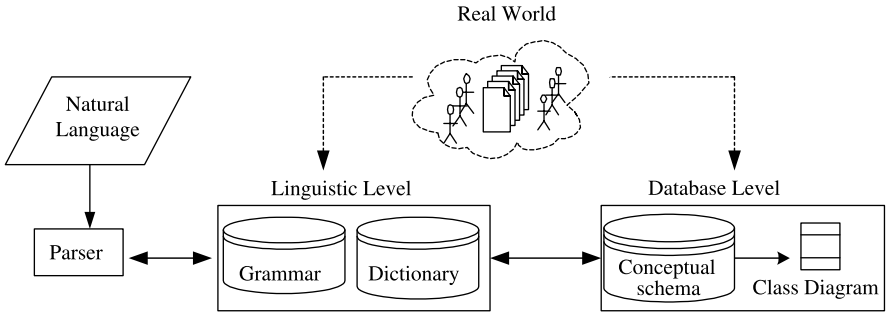

Fig. 1. The inter-relationship between the linguistic level and the database level.

world. The application of OO analysis can help us to identify how a class corresponds to an entity in the real world. Moreover, Moreno and van de Riet [15] have shown that conceptual modeling formalization of the basic constructs of English sentences can be mapped into OO conceptual model in a natural way.

In recent years, UML is becoming a widely accepted methodology for modeling and designing information systems. Due to its standardization, it is applied widely in the industrial and the academic societies. UML is a graphical notation for Object-Oriented system modeling to express requirement analysis and software design. The UML class diagram is an integrated part of UML and is used to design static data models. For the database schema that is designed by the UML class diagram, we have further extended the UML class diagram concepts to capture natural language semantics for database access in [24].

1.1. Objectives

In this paper, we present an approach to map the natural language queries with fuzzy semantics into SQL statements through the UML class diagram representations. Fuzzy queries allow users to express vague predicates, such as 'young' and 'good', to describe object of the real world more naturally. Additionally, fuzzy queries can possibly select a larger number of tuples in comparison to crisp ones. For example, even when crisp queries produce an empty result, the corresponding fuzzy queries can provide more possible answers for users. That makes queries with fuzzy terms more flexible than crisp queries. Based on [29–31], the theory of fuzzy set provides an application of the 'linguistic approach' for the modeling of natural language expressions.

Metais et al. [13] have pointed out that natural languages and the database conceptual schema can represent the conceptualization aspects of the real world. The inter-relationship between the linguistic level and the database level is shown in Fig. 1. In the linguistic level, natural language queries are analyzed by using the predefined grammar and an attached dictionary to reduce ambiguity and complexity. On the other side, in the database level, a schema can be regarded as the blueprint of the conceptual design of a database. Based on such observation, Owei et al. [18] proposed a concept-based query language that allows for the conceptual abstraction of database queries and exploits the rich semantics of semantic data models to ease and facilitate query formulation. In our work, for conceptual schemas organized by class diagrams, we will propose an approach for constructing a natural language interface.

The purpose of this paper is to explore the relationships between natural language constructs and the OO world. By referring to the database schema represented by the UML class diagram, our approach maps natural language constructs with fuzzy semantics into an extended class diagram representation, which will be further transformed into SQL statements for query executions according to the Structured Object Model (SOM) proposed in [9]. The approach handles natural language queries with fuzzy terms in combination with linguistic terms. That makes users issue both crisp and fuzzy query statements more conveniently.

1.2. Limitations

Many researchers believe that natural language interface (NLI) is an ideal means for user-system communication and have made a long effort to develop NLIs to database systems. However, as pointed out in [26], a frequent criticism concern of NLIs is that we cannot expect natural language interfaces to act appropriately for every input sentence. Therefore, we hope to limit our work in processing the following types of query sentences:

- Interrogative—e.g. "Does Smith supply monitors?"
- Imperative—e.g. "List all the suppliers."
- Declarative—e.g. "Smith supplies monitors." (Will be treated as a question.)

Moreover, users should be aware that the system might be unable to provide an answer if their expectations exceed the actual database capabilities, since the information stored in a database is just a precise world subset. If the system cannot make a decision to get answers due to some ambiguity, then users are asked to answer some questions to clarify the ambiguity.

2. Related work

2.1. Overview of natural language processing

Our prior research [23] has indicated that language-understanding process is commonly divided into three stages.