УДК 004.5, 004.652
DOI: 10.15827/0236-235X.122.291-294

Дата подачи статьи: 15.02.18
2018. Т. 31. № 2. С. 291–294

МЕТОД АВТОМАТИЗИРОВАННОГО ФОРМИРОВАНИЯ СЕМАНТИЧЕСКОЙ МОДЕЛИ БАЗЫ ДАННЫХ ДИАЛОГОВОЙ СИСТЕМЫ

Р.В. Посевкин [1], *аспирант, rus_posevkin@mail.ru*

[1] *Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), Кронверкский просп., 49, г. Санкт-Петербург, 197101, Россия*

Работа посвящена проблеме интеллектуального анализа содержимого БД для формирования семантической модели.

Для упрощения работы с программами на мобильных устройствах, таких как смартфоны и планшеты, активно внедряются голосовые ассистенты. По аналогии с голосовым ассистентом возможно внедрение диалоговой текстовой системы. Таким образом решается задача взаимодействия пользователя с программной системой использования более привычного естественного языка. Пользовательский интерфейс представляет собой систему программных решений, реализующих поиск, просмотр, получение и обработку информации из внешнего хранилища – БД. Естественно-языковой интерфейс является разновидностью пользовательского интерфейса, который принимает на вход и обрабатывает запросы на естественном языке, а также может использовать естественный язык для вывода найденной информации пользователю. Семантическая модель БД – важная компонента диалоговой системы. Данная модель описывает взаимосвязи и внутреннюю структуру БД. Формирование семантической модели БД вручную приводит к существенному увеличению временных и трудозатрат, стоимости разработки программной системы.

Цель автора данной работы – автоматизация процесса формирования семантической модели БД диалоговой системы. Предлагаемый метод состоит в применении ряда подходов, позволяющих в автоматизированном режиме формировать семантическую модель существующей БД.

Используя тезаурус предметной области, можно определить семантику, в значительной степени решив проблему многозначности при интерпретации текста. Применение паттернов позволяет выявить связи внутри БД. Анализ содержимого полей БД дает возможность определить характер и семантику хранимых данных, а указание локали – сократить время, необходимое для анализа содержимого БД.

Ключевые слова: *обработка естественного языка, семантическая модель, пользовательский интерфейс, БД, диалоговая система.*

В настоящее время существует большое количество операционных систем и прикладного ПО. Каждая из программ обладает специфичным интерфейсом взаимодействия, что приводит к увеличению времени, необходимого для обучения пользователя работе с системой. Для упрощения работы с программами на мобильных устройствах, таких как смартфоны и планшеты, активно внедряются голосовые ассистенты. По аналогии с голосовым ассистентом возможно внедрение диалоговой текстовой системы, представляющей собой интерактивную программную систему человеко-машинного взаимодействия, где пользователь может формировать запросы на естественном языке. Таким образом решается задача взаимодействия пользователя с программной системой использования более привычного естественного языка [1].

Пользовательский интерфейс представляет собой систему программных решений, реализующих поиск, просмотр, получение и обработку информации из внешнего хранилища – БД. Естественно-языковой интерфейс является разновидностью пользовательского интерфейса, который принимает на вход и обрабатывает запросы на естественном языке, а также может использовать естественный язык для вывода найденной информации пользователю [2]. Существуют различные подходы к разработке естественно-языковых интерфейсов. Например, в основе проекта NaLIR [3] лежат формирование дерева зависимостей, а также использование эвристик и правил в процессе разбора естественно-языкового запроса, в то время как в проекте Sqlizer [4] используются методы машинного обучения.

Семантическая модель представляет собой одну из важнейших компонент диалоговой системы. Данная модель описывает взаимосвязи и внутреннюю структуру БД [5]. Особенностями русского языка являются его гибкость и многозначность, что значительно затрудняет интерпретацию запросов. Таким образом, семантическая модель БД может быть использована для решения задач, связанных с разрешением неоднозначности естественного языка [6].

Процесс обработки пользовательского запроса на естественном языке состоит из последовательного выполнения морфологического, синтаксического и семантического анализа [7]. Первыми осуществляются морфологический и морфемный анализы пользовательского запроса. В рамках морфологического анализа определяются падеж, склонение, часть речи. При морфемном анализе каждое слово разбивается на отдельные морфемы: приставка, корень, суффикс, окончание.

Во время синтаксического анализа выделяются синтаксические связи внутри предложения – глав-

ные и второстепенные члены предложения, тип предложения. На данном этапе используются синтаксические и лексические правила анализируемого языка, а также информация, полученная на этапе морфологического анализа [8].

Следующий шаг обработки естественно-языкового запроса – построение семантического представления. Семантическое представление естественно-языкового запроса пользователя строится на основе семантической модели БД.

Формирование семантической модели БД. В семантической модели описываются сущности, информация о которых содержится в БД. Также модель включает в себя отношения между сущностями, которые аналогичны связям в диаграммах сущность–связь (ER-диаграммы). При разработке диалоговой системы перед разработчиком зачастую стоит задача реализации естественно-языкового пользовательского интерфейса к уже существующей и наполненной данными БД. В таком случае ручное формирование семантической модели БД приводит к существенному увеличению количества временных и трудозатрат и, как следствие, стоимости разработки программной системы.

Для решения проблемы требуется разработать механизм, помогающий в автоматизированном режиме формировать семантическую модель существующей БД. Формирование семантической модели происходит в следующей последовательности (рис. 1):

– извлекаются названия всех таблиц БД;
– для каждой таблицы извлекаются все поля;
– на основе названий таблиц и их полей делается предположение о содержимом, определяются внутренние связи;
– в ручном режиме возможны уточнение и расширение информации, полученной на предыдущем этапе.

Одной из проблем, препятствующих однозначному корректному определению содержимого БД, является многозначность естественного языка. Поля таблиц с одним и тем же содержимым могут иметь множество различных названий в зависимости от того, кто их задавал.

Определение внутренних связей БД. Для определения внутренних связей между таблицами возможно применение ряда паттернов, основанных на

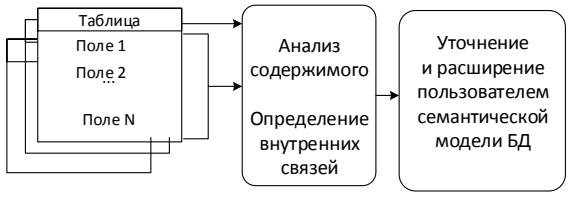

Рис. 1. Процесс формирования семантической модели БД

Fig. 1. Database semantic model construction process

именовании таблиц и полей БД. Одним из подобных паттернов является связь внешнего ключа и таблицы БД вида *id_*[*tableName*] или [*tableName*]*_id*, где [*tableName*] – имя таблицы БД, с которой связан внешний ключ.

Рассмотрим конкретный пример (рис. 2). Имеются таблица *employee* – переменная отношения R_2 и таблица *department* – переменная отношения R_1. При этом поле *id_department* таблицы *employee* является внешним ключом (*FK*), значения которого должны совпадать со значениями потенциального ключа (*CK*) переменной отношения R_1. В роли такого потенциального ключа выступает первичный ключ – поле *id* таблицы *department.* При этом выполняются следующие условия:

– в переменной отношения R_1 имеется потенциальный ключ *CK*, такой, что *FK* и *CK* совпадают с точностью до переименования атрибутов;
– в любой момент времени каждое значение *FK* в текущем значении R_2 идентично значению *CK* в некотором кортеже в текущем значении R_1.

Таким образом, с применением паттерна вида *id_*[*tableName*] обнаружена взаимосвязь родительского R_1 (*department*) и дочернего R_2 (*employee*) отношений.

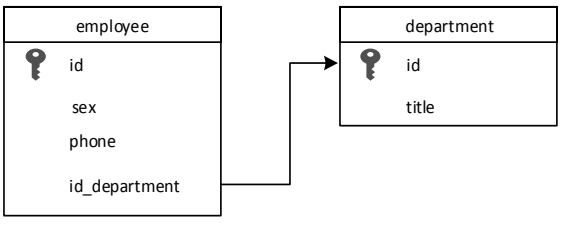

Рис. 2. Связь таблиц БД

Fig. 2. Tables connections in a database

Определение семантики и типа связи между сущностями БД с использованием тезауруса. Одним из вариантов решения проблемы определения семантики данных, хранящихся в том или ином поле таблицы БД, является применение тезауруса предметной области. Тезаурус в общем виде представляет собой некий словарь, включающий в себя понятия, определения и термины специальной области знаний [9]. В рамках этой области знаний хранится информация в БД. Тезаурус также может включать семантические отношения между лексическими единицами, такие как синонимы, антонимы и т.п. Тезаурус позволяет выявить смысл не только с помощью определения, но и посредством соотнесения слова с другими понятиями и их группами.

Тезаурус необходимо сформировать заранее. В качестве подобного тезауруса могут быть использованы общие решения, например EuroWordNet [10], либо собственный тезаурус. Собственный тезаурус может быть самостоятельно сформирован в автоматическом режиме, например, на основе

технической документации проекта или иных метаданных.

В рассмотренном выше примере выявлена связь между сущностями *сотрудник (employee)* и *отдел (department)*, расположенными в разных таблицах БД. При этом остается неизвестным вид связи между выделенными сущностями. Эту проблему можно решить, если предоставить пользователю возможность задать вид связи вручную после автоматического определения сущностей.

В качестве альтернативного подхода создается пользовательский тезаурус, включающий в себя виды связей между сущностями, хранящимися в БД. Подобный тезаурус разрабатывается в рамках организации, занимающейся сопровождением и наполнением БД. В данном случае связь между сущностями *employee* и *department* представлена в виде следующего триплета: < *employee, работает в, department* >.

При этом, дополнительно применяя информацию из общих тезаурусов, можно определить семантику сущностей, хранящихся в рассматриваемых таблицах БД. В результате получается отношение, заданное следующим триплетом: < *сотрудник, работает в, отдел* >.

Определение семантики на основе анализа содержимого. Одним из вариантов определения семантики поля в автоматическом режиме является анализ содержимого этого поля. В качестве примера можно привести адрес электронной почты, который представляет собой запись в виде строки, состоящей из двух частей, разделенных символом @: [*prefix*]@[*postfix*], где [*prefix*] представляет собой текстовую строку, а [*postfix*] – многоуровневое доменное имя. В соответствии с RFC 5322 [11] проверка на наличие адреса электронной почты может быть осуществлена с помощью регулярного выражения, схема формирования которого представлена на рисунке (см. http://www.swsys.ru/uploaded/image/2018_2/2018-2-dop/16.jpg). При этом в связи с распространением национальных доменов первого уровня [*prefix*] и [*postfix*] могут быть представлены не только на английском языке и данное регулярное выражение может быть расширено для поддержки символов национальных алфавитов.

Еще одним подходом к анализу содержимого поля БД является использование различных текстовых корпусов. Например, подобным образом возможно подтверждение гипотезы по нахождению связки Фамилия–Имя–Отчество с применением соответствующего текстового корпуса. В случае русского языка процедура проверки предположения наличия Ф.И.О. в поле может быть инициирована при наличии в поле трех слов, начинающихся с заглавной буквы.

В качестве дополнительной возможности улучшения качества анализа данных возможно указание локали данных перед этапом анализа содержимого БД. Это позволит еще и сократить время обработки данных, так как в этом случае будут использоваться специфичные для указанных языков правила, тезаурусы и текстовые корпуса.

Результаты экспериментального исследования. Проведено экспериментальное исследование предложенного метода автоматизированного формирования семантической модели БД диалоговой системы. Для эксперимента создана диалоговая система, для которой в автоматизированном режиме была сформирована семантическая модель. Далее несколькими добровольцами на естественном языке были сформированы вопросы к экспериментальной диалоговой системе. Таким образом сформировано $|D_{rel}| = 130$ вопросов.

По результатам взаимодействия с диалоговой системой вручную были оценены корректность сформированных SQL-запросов к БД и релевантность полученного ответа. Синтаксически корректно сформированными оказались $|D_{retr}| = 84$ SQL-запроса. При этом релевантный ответ был получен в $|D_{rel} \cap D_{retr}| = 75$ случаях. Последующий анализ ситуаций, в которых не удалось построить корректный SQL-запрос или получить релевантный ответ, показал, что использование пользовательского тезауруса, содержащего информацию об аббревиатурах и сокращениях, позволил бы улучшить полученный результат. В итоге точность извлечения $Pr = |D_{rel} \cap D_{retr}| / |D_{retr}| = 0.89$, полнота $Re = |D_{rel} \cap D_{retr}| / |D_{rel}| = 0.58$, комбинированная F-метрика $2PrRe / (Pr + Re) = 0.70$.

Заключение

Семантическая модель БД является важной компонентой диалоговой системы. Формирование семантической модели БД вручную приводит к существенному увеличению временных и трудозатрат, стоимости разработки программной системы.

В работе предложен ряд подходов, позволяющих в автоматизированном режиме формировать семантическую модель существующей БД. Использование тезауруса предметной области дает возможность определить семантику, в значительной степени решив проблему многозначности при интерпретации текста. Применение паттернов позволяет выявить связи внутри БД, интеллектуальный анализ содержимого полей БД – определить характер хранящихся данных, а указание локали – сократить время, требуемое для анализа содержимого БД. Предложенный подход устраняет недостатки существующих решений [12], так как предполагает наличие информации о внутренней структуре БД, что помогает построить более точный SQL-запрос.

Литература

1. Llopis M., Ferrández A. How to make a natural language interface to query databases accessible to everyone: An example. Computer Standards & Interfaces. 2013, vol. 35, i. 5, pp. 470–481.

2. Zhou L., Mohammed A.S., Zhang D. Mobile personal information management agent: Supporting natural language interface and application integration. Information Processing & Management. 2012, vol. 48, i. 1, pp. 23–31.

3. Li F., Jagadish H.V. NaLIR: an interactive natural language interface for querying relational databases. Proc. ACM SIGMOD, 2014, pp. 709–712.

4. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality. Proc. 26th Intern. Conf. on Neural Inform. Processing Syst., 2013, pp. 3111–3119.

5. Giordani A., Moschitti A. Semantic mapping between natural language questions and SQL queries via syntactic pairing. Intern. Conf. on Application of Natural Language to Inform. Syst. Springer Berlin Heidelberg, 2009, pp. 207–221.

6. Сулейманов Д.Ш. Двухуровневый лингвистический процессор ответных текстов на естественном языке // Междунар. науч.-технич. конф. OSTIS-2011: сб. тр. Минск: Изд-во БГУИР, 2011. С. 311–322.

7. Posevkin R., Bessmertny I. Multilanguage natural user interface to database. Proc. 10th Intern. Conf. IEEE AICT-2016. 2016, pp. 304–306.

8. Bhadgale A.M., Gavas S.R., Goyal P.R. Natural language to SQL conversion system. IJCSEITR. 2013, vol. 3, i. 2, pp. 161–166.

9. Браславский П.И. Тезаурус для расширения запросов к машинам поиска Интернета: структура и функции // Диалог: тр. Междунар. конф. 2003. С. 95–100.

10. Vossen P. EuroWordNet: a multilingual database for information retrieval. Proc. DELOS Workshop on Cross-language Information Retrieval, 1997, pp. 5–7.

11. Resnick P.W. Internet message format. 2008. URL: http://www.ietf.org/rfc/rfc5322.txt (дата обращения: 14.01.2018).

12. Brad F., Iacob R., Hosu I., Rebedea T. Dataset for a Neural Natural Language Interface for Databases (NNLIDB). Proc. 8th IJCNLP, 2017, pp. 906–914.

Software & Systems
DOI: 10.15827/0236-235X.122.291-294

Received 15.02.18
2018, vol. 31, no. 2, pp. 291–294

A METHOD OF AUTOMATED DEVELOPMENT OF THE SEMANTIC DIALOGUE SYSTEM DATABASE MODEL

R.V. Posevkin [1], *Postgraduate Student, rus_posevkin@mail.ru*

[1] *The National Research University of Information Technologies, Mechanics and Optics, Kronverksky Ave. 49, St. Petersburg, 197101, Russian Federation*

Abstract. The paper considers the problem of intellectual database content analysis for creating a semantic database model.

Voice assistants are created to simplify interaction with mobile devices such as smartphones and tablets. A text dialogue system is an analogue of this approach. As a result, user can interact with software using natural language. User interface is a set of software solutions that helps to search, review, obtain and process information from a database that is external storage. Natural language interface is a sort of user interface that accepts and processes natural language queries. This interface can also use natural language in output to show found information to a user. A semantic database model is an important part of the dialogue system. This model includes interconnections and internal structure of a database. Manual creating of a semantic database model significantly increases time and labour costs, as well as development cost of a software system.

The main purpose of the article is automation of development of a dialogue system semantic database model. The proposed method uses a set of approaches to automated creating of an existing database semantic model.

An object domain thesaurus helps to define semantics and solve the problem of polysemy in text processing. Patterns helps to extract interconnections in a database. The analysis of database field content allows determining data semantic and nature. Locale indication allows decreasing the time for a database content analysis.

Keywords: natural language processing, semantic model, user interface, database, dialogue system.

References

1. Llopis M., Ferrández A. How to make a natural language interface to query databases accessible to everyone: An example. *Computer Standards & Interfaces*. 2013, vol. 35, i. 5, pp. 470–481.

2. Zhou L., Mohammed A.S., Zhang D. Mobile personal information management agent: Supporting natural language interface and application integration. *Information Processing & Management*. 2012, vol. 48, i. 1, pp. 23–31.

3. Li F., Jagadish H.V. NaLIR: an interactive natural language interface for querying relational databases. *Proc. 2014 ACM SIGMOD Int. Conf. on Management of Data*. 2014, pp. 709–712.

4. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality. *Proc. 26th Int. Conf. on Neural Information Processing Systems*. 2013, pp. 3111–3119.

5. Giordani A., Moschitti A. Semantic mapping between natural language questions and SQL queries via syntactic pairing. *Int. Conf. on Application of Natural Language to Information Systems*. Springer Berlin Heidelberg Publ., 2009, pp. 207–221.

6. Suleymanov D.Sh. Dual-layer linguistic processor of the answering natural language texts. *Sb. tr. Mezhdunar. nauch.-tekhnich. konf. OSTIS-2011* [Proc. Int. Sci. and Tech. Conf. OSTIS-2011]. Minsk, BGUIR Publ., 2011, pp. 311–322 (in Russ.).

7. Posevkin R., Bessmertny I. Multilanguage natural user interface to database. *Proc. 10th Int. Conf. IEEE AICT-2016*. 2016, pp. 304–306.

8. Bhadgale A.M., Gavas S.R., Goyal P.R. Natural language to SQL conversion system. *IJCSEITR*. 2013, vol. 3, i. 2, pp. 161–166.

9. Braslavsky P.I. Thesaurus for extension queries to search machine: structure and functions. *Kompyuternaya lingvistika i intellektualnye tekhnologii. Tr. mezhdunar. konf. Dialog* [Computer Linguistics and Intelligent Technologies. Proc. Int. Conf. Dialog]. 2003, pp. 95–100 (in Russ.).

10. Vossen P. EuroWordNet: a multilingual database for information retrieval. *Proc. DELOS Workshop on Cross-Language Information Retrieval*. 1997, pp. 5–7.

11. Resnick P.W. *Internet message format*. 2008. Available at: http://www.ietf.org/rfc/rfc5322.txt (accessed January 14, 2018).

12. Brad F., Iacob R., Hosu I., Rebedea T. Dataset for a Neural Natural Language Interface for Databases (NNLIDB). *Proc. 8th Int. Joint Conf. on Natural Language Processing*. 2017, pp. 906–914.