Fomichov, V.A. The Theory of K-representations as a Flexible Tool for Designing Natural-Language Interfaces of Recommender Systems // Kurt J. Engemann and George E. Lasker (Eds.) Advances in Decision Technology and Intelligent Information Systems, Vol. XIII.
The International Institute for Advanced Studies in Systems Research and Cybernetics (IIAS), Tecumseh, Ontario, Canada, 2012, p. 6-10, ISBN 978-1-897546-67-3

The Theory of K-representations as a Flexible Tool for Designing Natural-Language Interfaces of Recommender Systems

Vladimir A. Fomichov

Department of Innovations and Business
in the Sphere of Informational Technologies
Faculty of Business Informatics
National Research University Higher School of Economics
Kirpichnaya str. 33, 105187 Moscow, Russia

vfomichov@hse.ru and vfomichov@gmail.com

Abstract

The paper describes a new method of constructing recommender systems with natural-language interface. This method is based on the theory of K-representations (knowledge representations) - a new theory of designing semantic-syntactic analyzers of natural language texts with the broad use of formal means for representing input, intermediary, and output data. The current version of the theory is set forth in a monograph published by Springer in 2010. The stated approach is implemented in the programming environment PHP + MySQL: an experimental recommender system has been developed.

Keywords: recommender system with natural-language interface; semantic representation of input request; theory of K-representations; SK-languages; algorithm of semantic-syntactic analysis

Introduction

Since the end of the 1990s, a new branch of e-commerce has been quickly developing, this branch is called Recommender Systems (RecS). The software applications of this class are intended for consulting the end users of the Internet with the aim of helping them to take the decisions about the choice of goods or/and services. An experiment carried out in USA showed that some 80 percents of the RecS users prefer to work with a natural language interface to RecS but not with numerous displayed menus (the English language has been used in this experiment) (Chai et al, 2002). That is why the key role in the functioning of many RecS is played by the interaction with the users by means of Natural Language (NL) – English, Russian, German, etc.

In the monograph (Fomichov, 2010) a new theory of designing semantic-syntactic analyzers of NL-texts with the use of formal means for representing input, intermediary, and output data is proposed. This theory, called the theory of K-representations, can be interpreted as powerful and flexible tool of designing the NL-interfaces of RecS. Let’s consider the structure of this theory.

The *first basic constituent* of the theory of K-representations is the theory of SK-languages (standard knowledge languages), stated, in particular, in (Fomichov, 2010). The kernel of the theory of SK-languages is a mathematical model describing a system of such 10 partial operations on structured meanings (SMs) of natural language texts (NL-texts) that, using primitive conceptual items as "blocks", we are able to build SMs of arbitrary NL-texts (including articles, textbooks, etc.) and arbitrary pieces of knowledge about the world. The analysis of the scientific literature on artificial intelligence theory, mathematical and computational linguistics shows that today the class of SK-languages opens the broadest prospects for building semantic representations (SRs) of NL-texts (i.e., for representing meanings of NL-texts in a formal way).

The *second basic constituent* of the theory of K-representations is a broadly applicable mathematical model of a linguistic database. The model describes the frames expressing the necessary conditions of the existence of semantic relations, in particular, in the word combinations of the following kinds: “Verbal form (verb, participle, gerund) + Preposition + Noun”, “Verbal form + Noun”, “Noun1 + Preposition + Noun2”, “Noun1+ Noun2”, “Number designation + Noun”, “Attribute + Noun”, “Interrogative word + Verb”.

The *third basic constituent* of the theory of K-representations is a complex, strongly structured algorithm carrying out semantic-syntactic analysis of texts from some practically interesting sublanguages of NL. The algorithm *SemSynt1* transforms a NL-text into its semantic representation being an expression of a certain SK-language (Fomichov 2010). An important feature of this algorithm is that it doesn’t construct any syntactic representation of the inputted NL-text but directly finds semantic relations between text units. The other distinguished feature is that a complicated algorithm is described with the help of formal means, that is why it is problem independent and doesn’t depend on a programming system. The algorithm is implemented in the Web programming language PHP. It will be shown in the paper that the theory of K-representations opens broad prospects for designing NL-interfaces of recommender systems.

A new method of designing recommender systems with a natural language interface

A new method of formalizing the design of RecS with a semantics-oriented NL-interface has been elaborated. Its basic assumptions are the following principles of transforming a normalized NL-request of the user of a RecS into an SQL-request. Firstly, a request is transformed into a semantic representation (SR), where SR is built with the help of a broadly applicable, domain independent, and program independent formalism. On the next stage, a SR of the request is transformed into an SQL-request. During these stages, the clarifying questions can be forwarded to the user. For making easier and less expensive the design of a NL-interface, it is proposed to introduce the requests including only the nouns, attributes, the word “not”, the comparative forms of attributes (“not more expensive than”), and the conjunctions “and”, “or”. The proposed method includes the following stages of design:

1. The development of a problem-oriented mathematical model of the system of primary conceptual items used by the NL-interface of a recommender system (RecS), proceeding from the first basic mathematical model of the theory of K-representations determining the class of formal objects called *conceptual bases* .

2. The construction of a mathematical model of the variety of conceptual structures corresponding to the normalized NL-requests of the users of a RecS with the help of SK-languages, proposed by the theory of K-representations.
3. The development of a mathematical model of a linguistic database (LDB), i.e. database containing the information used by the algorithms of semantic-syntactic analysis for transforming the input requests into their semantic representations (see Chap. 7 of (Fomichov, 2010)).
4. The creation of an algorithm transforming a normalized NL-request of the user of a RecS into its SR being its K-representation, that is, being an expression of a certain SK-language.
5. The development of an algorithm transforming a K-representation of an input request into an SQL-expression for (a) organization (on the next stage) the interaction with a database containing the information about goods (or services) and (b) formulating the recommendations for the user.
6. The elaboration of the structure of a relational database for storing morphological and semantic-syntactic information associated with the lexical items.
7. The development of a computer program for transforming a normalized NL-request of the user of a RecS into an SQL-request.
8. The design and program implementation of an algorithm organizing a dialogue with the user of a RecS.

Mathematical representation of conceptual structures corresponding to the normalized natural language requests of the users of a recommender system

The starting point for the first step of the stated method is the definition of a new class of formal objects called conceptual bases. This definition is provided by the theory of K-representations (see Chapter 3 of (Fomichov, 2010). To define an arbitrary conceptual basis (c.b.) is equivalent to determining a certain 15-tuple of the form âèäà (c_1, c_2, …, c_{15}) with the components being mainly the countable or finite sets of symbols or the distinguished elements of these sets. As a result of considering a number of additional definitions, the class of so called problem-oriented conceptual bases (p.o.c.b.) is introduced. The series of all these definitions is interpreted as a problem-oriented mathematical model of the system of primary conceptual items used by the NL-interface of a recommender system (RecS).

In order to describe in a mathematical way the conceptual structures corresponding to the initial NL-requests of the users of a RecS, an arbitrary p.o.c.b. *Probs* and an arbitrary variable *var* are associated with a formal language *Linf(B,var)*, where *B* = *B(Probs)* is the conceptual basis being the first component of the p.o.c.b. *Probs*.

Example. It is possible to define a p.o.c.b. *Probs* in such a way that *y*$_1$ *Î V(B(Probs))* and *Linf(B(Probs),y*$_1$*)* includes the string *(Less1(Price(y1),35000/EUR)Ù Color(y1,(dark-green Ú dark-blue)) Ù Country-of-assembly(y*$_1$*,(Germany Ú Belgium))).*

Most often, a request consists of two parts. The part 1 shortly designates the object of interest of the user (e.g., “a German car”). The part 2 lists additional requirements to be satisfied by any object of interest. That is why it is proposed to construct *a primary semantic image (PSI)* of a request in the form <*Semrepr1*, *Semrepr2*>, where *Semrepr1* is a semantic representation (SR) of a short description of the object of interest, and *Semrepr2* is a SR of a fragment enumerating additional requirements to be satisfied by any object of interest (for example, “not older than 5 years, the color dark-green or dark-blue”).

Example. Let the request 1 **= “**German hatchbacks not older than 5 years, the color dark-green or dark-blue”, not more expensive than 35.000 euros”. Then a p.o.c.b. *Probs* can be determined in such a way that a PSI of the request 1 will be the expression of the form <*Semrepr1, Semrepr2*>, where *Semrepr1=all car1 * (Country-of-assembly, Germany)(Form-of-car, hatchback), Semrep2=(¬Greater2(Age(*y_1*), 5/year) Ù Color(*y_1*,(dark-green Ú dark-blue)) Ù ¬Greater1(Price(y1), 35000/EUR)).*

The values of the parameters *Country-of-assembly* and *Form-of-car* are indicated in the substring *Semrepr1* not in the same form as the values of the parameters *Age, Color, and Price* are represented in the substring *Semrepr2*. The analysis of the expressive mechanisms of SK-languages has permitted to propose such form of *a primary semantic representation (PSR)* of the inputted natural language request of a RecS user that this form allows for representing in the same way the values of all parameters of the objects of interest, and it is convenient for the processing of a request on the next stage. If *semrequest* is a PSI of a request, then a SR of the request *semrequest* is a value of a special mapping *Secondary-form.*

Example. Let the request 1 **= “**German hatchbacks not older than 5 years, the color dark-green or dark-blue, not more expensive than 35.000 euros”, and *semrequest* is a primary semantic image of the request 1 being the string of the form <*Semrepr1, Semrepr2*>. Then a PSR of the considered request *Secondary-form(semrequest)* is the string of the form

Object-of-interest (request$_1$*, all car1 *(Element, S*$_1$*), Description1(arbitrary*
car1(Element, S*$_1$*):y*$_1$*, (Country-of-assembly (y*$_1$*, Germany) Ù*
Form-of-car(y$_1$*, hatchback) Ù ¬Greater1(Age(y*$_1$*), 5/year) Ù Color(y*$_1$*, (dark-green Ú dark-*
blue)) Ù ¬ Greater2 (Price(y$_1$*), 35000/EUR)).*

In order to transform the requests inputted by the users of a RecS into their SR, a mathematical model of a linguistic database (LDB) has been developed. The model defines a new class of formal objects called the *problem-oriented linguistic bases*. According to this model, the main components of a LDB being a part of a RecS are a lexico-semantic dictionary and a dictionary of prepositional frames. The structure of these dictionaries is proposed in Chapter 7 of (Fomichov, 2010). Two algorithms have been developed. The first algorithm transforms a normalized NL-request of the user (a request in normalized Russian language) into its semantic representation being a K-representation of the request (i.e., an expression of the SK-language in a certain conceptual basis). The second algorithm transforms a K-representation of the request into an SQL-expression for organizing the interaction with the database about the goods or services and for formulating the recommendations for the user.

For implementing the algorithms, a structure of a MySQL database has been elaborated, this structure is presented in the form of an ER-model. In order to connect the interpreter with the data about the goods and their characteristics, the intermediary tables are used, where the values of properties in NL and their semantic representations are linked with the corresponding fields associated with goods. Since the proposed structure is very flexible, an SQL-request may have, in particular, the following form:

*select * from country, country_kategories,_tovar, kategories where country.country_name = 'Germany' and country.country_id = country_kategories.country_id and tovar.id_kategorii = kategories.id_kategories and kategories.id_pred_kategorii = country_kategories.id_kategories* .

A request may mention more than one property (attribute), for instance, it may be the string

*Object-of-interest (request$_2$, all car1 * (Element, S$_2$),*
*Dsecription1 (arbitrary car1 * (Element, S$_2$) :y$_2$, ((Country-of-assembly (y$_2$, France) Ù Form-of-car (y$_2$, sedan) Ù ¬ Greater1(Price(y$_2$), 14000/USD)))).*

This request will be associated with the SQL-expression *select * from auto where auto.country=’France’ and cars.body.type=’sedan’ and cars.price<=14000.*

Program implementation and experimental results

For implementing the method stated above, a RecS in the programming environment PHP + MySQL has been developed (Pravikov and Fomichov, 2010). The testing of its first version has shown that an average request including two characteristics of an object of interest is performed during 0.09 sec, and the number of SQL server calls is 12. In order to study the work of the developed scripts, these scripts have been incorporated into an existing Web site with a real database about the cars. During the experiment, on the basis of the rating LiveInternet, the growth of the number of visitors from 20 to 28 percents has been observed in comparison with the similar time interval one week before. Besides, the average time for visiting this site became 37 percents higher than before.

Conclusion

The analysis of the expressive power of SK-languages shows that the theory of K-representations can be used as a powerful and flexible tool for designing arbitrary semantics-oriented natural language interfaces of recommender systems.

References

1. Chai, J., Horvath, V., Nicolov, N., Stys, M., Kambhatla, N., Zadrozny, W., Melville, P. (2002) Natural Language Assistant – A Dialog System for Online Product Recommendation; AI Magazine, V. 23, No. 2 (pp. 63-76)

2. Fomichov, V.A. (2010); Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Series: IFSR International Series on Systems Science and Engineering, Vol. 27.; Springer, New York, Dordrecht, Heidelberg, London (354 pp)

Pravikov, A.A. and Fomichov, V.A. (2010); Development of a Recommender System with Natural Language Interface on the Basis of Mathematical Models of Semantic Objects; Business-Informatics (Moscow), No. 4 (14), in Russian (pp. 3-11)