Object Oriented Modelling in Information Systems Based on Related Text Data

Kolyo Onkov,

Agricultural University, Department of Computer Science and Statistics, Mendeleev 12, 4000 Plovdiv, Bulgaria
kolyoonkov@yahoo.com

Abstract. Specialized applied fields in natural sciences – medicine, biology, chemistry etc. require building and exploring of information systems based on related text forms (words, phrases). These forms represent expert information and knowledge. The paper discusses the integration of two basic approaches – relational for structuring complex related texts and object oriented for data analysis. This conception is implemented for building of information system "Crop protection" in Bulgaria based on the complex relationships between biological (crops, pests) and chemical (pesticides) terms in textual form. Analogy exists between class objects in biology, chemistry and class objects and instances of object oriented programming. That fact is essential for building flexible models and software for data analysis in the information system. The presented example shows the potential of object oriented modelling to define and resolve complex tasks concerning effective pesticides use.

Keywords: expert data, natural sciences, key words, text data retrieval, relational database, object oriented modelling, crop protection

1 Introduction

Text is the most common form for presenting expert information and knowledge. Paper guides and reference books using knowledge from natural sciences – medicine, biology, physics, chemistry etc., aim to impose the state rules and regulations in order to keep proofs of decisions made and to share and disseminate the useful information. As being heavily used by specialists it is difficult when the volume of text data is large and there are complex relationships between basic terms in textual form, for example the names of drugs and diseases. The following basic problems arise in the process of transformation text data containing related key words and phrases into intelligent computer based system: a) to retrieve related key words and phrases and to define relationships between them; b) to store text data and relationships in easy accessible database; c) to develop software for database analysis to meet information and knowledge requirements determined by specialists in the applied fields.

This paper reveals one conception for building of information systems based on related text data through integration of two approaches – relational for data structuring and object oriented for data modelling and analysis. This conception is implemented for the development of Bulgarian information system "Crop protection".

2 Conception of Information System Development

The conceptual framework of building and exploring of information system based on related text data is presented in figure 1.

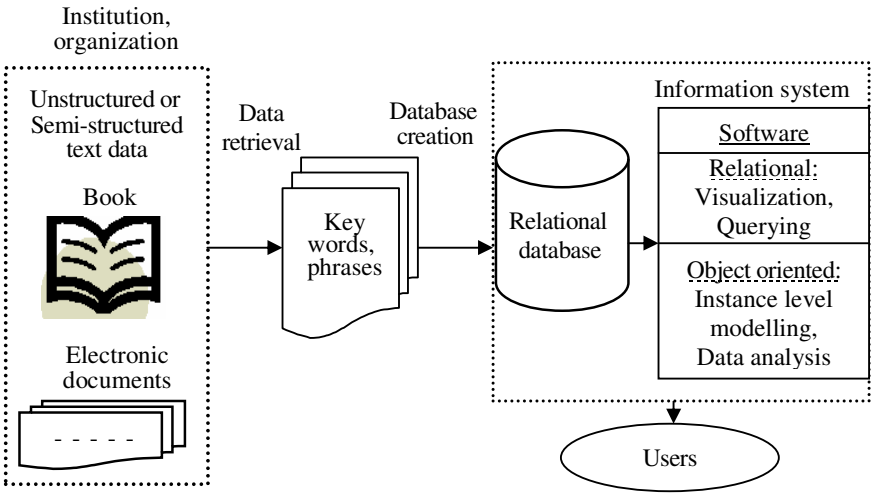


Figure 1: Conception of building and exploring of information system based on related text objects

In many cases paper or electronic guides and reference books in pharmacology, veterinary, agronomy etc present knowledge from natural sciences in the form of semi-structured or unstructured text data [1], [14]. The core of these expert texts is a set of contextually related words and phrases (terms in the fields). The retrieval of "key words" and relationships among them is usually done through text mining methods and algorithms [2], [3], [4], [5], [6]. The process of verification and correction [7], [8] is needed to guarantee storing correct expert data in the database.

Figure 2 presents information model for treatment of subjects which is common for several natural sciences: medicine, biology, chemistry and pharmacology. The relationships between terms (key words, elements of the sets) exist in both sequences: (M_1 _ M_2 _M_3) and (M_3 _ M_2 _M_1). Coding of key words is applied for building entity-relationships database model [9]. Ontology-based extraction and structuring [10], conceptual modelling of XML data [11] and relational system between two databases [12] are developed for structuring this type of data.

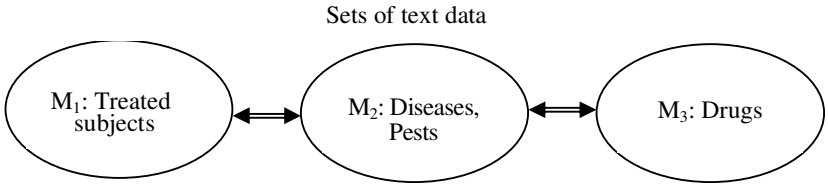


Figure 2: Information model for treatment of subjects

Relational software provides easy access to data while object oriented approach is the proper paradigm for building flexible models. The objects classifications in natural sciences are usual form of knowledge presentation. That is additional and important argument for building object oriented software for data analysis.

3 Object Oriented Modelling in Bulgarian Information System "Crop Protection"

The Bulgarian ministry of agriculture prepares annually a reference book [1] presenting pesticides permitted for use in the country (figure 3). The book content is based on the state laws and knowledge of national experts from the field.

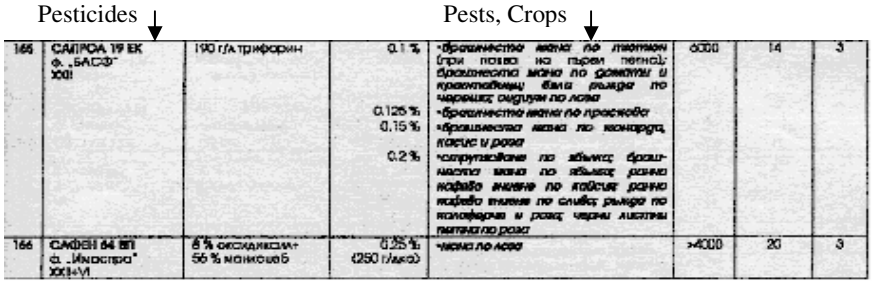


Figure 3: Scan copy of a fragment from one sheet of the reference book

The conception from the previous chapter is applied to transform the reference book into information system "Crop protection". The relationships between basic biological (crops, pests) and chemical (pesticides) terms are implemented.

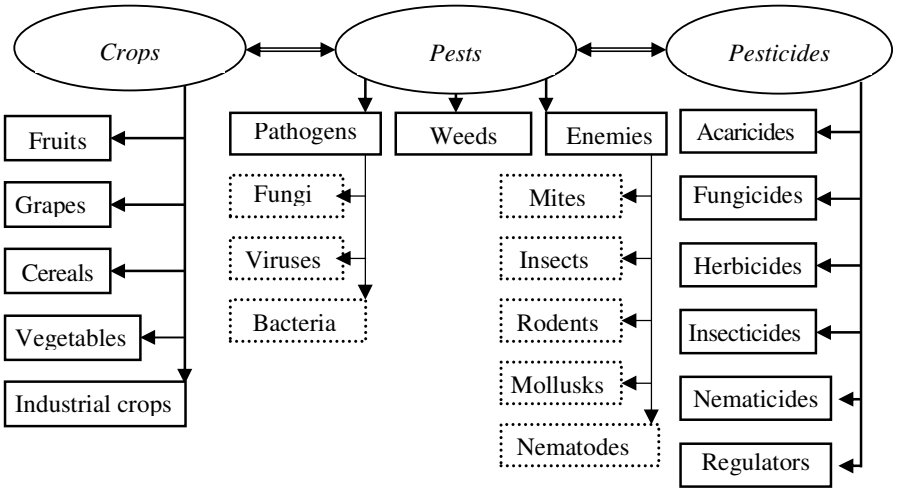
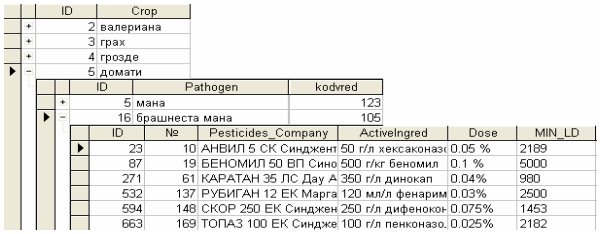


Figure 4: Information model for crop protection

The information model for crop protection (figure 4) consists of three hierarchical structures. The nodes of the model are the sets “Crops”, “Pests” and “Pesticides”. Each node is built by class objects corresponding to the biological and chemical classification. Each class contains text data related with the text data from other class.

Key words and phrases belonging to different classes and subclasses are extracted from the reference book and coded by natural numbers. Then the codes are used as primary keys of table schemes in the relational database. Figure 5 illustrates data access and visualization through relationships in the sequence “crops–pests (pathogens) –pesticides”. The quantitative and qualitative characteristics of each pesticide – active ingredient, dose and minimum lethal dose (LD) are also presented.



ID	Crop
2	валериана
3	грах
4	грозде
5	домати

ID	Pathogen	kodvred
5	мана	123
16	брашнеста мана	105

ID	№	Pesticides_Company	ActiveIngred	Dose	MIN_LD
23	10	АНВИЛ 5 СК Синджент	50 г/л хексаконазо	0.05 %	2189
87	19	БЕНОМИЛ 50 ВП Сино	500 г/кг беномил	0.1 %	5000
271	61	КАРАТАН 35 ЛС Дау А	350 г/л динокап	0.04%	980
532	137	РУБИГАН 12 ЕК Марга	120 мл/л фенарим	0.03%	2500
594	148	СКОР 250 ЕК Синджен	250 г/л дифеноко	0.075%	1453
663	169	ТОПАЗ 100 ЕК Синдже	100 г/л пенконазо	0.025%	2182

Figure 5: Relational database – data access and visualization

Probably the easy access to data is enough for farmers and agronomists, but scientists and experts in biology, chemistry and phytopharmacy usually have to solve more complex tasks which require deep analysis of related text data. The complexity of data analysis descends from the close properties and at the same time variations of subjects from definite class, as well as variations among classes. That is the main reason to create flexible data models by using of object oriented approach. The coding of key words which is corresponding to the subjects from definite class gives opportunity for transition from relational to object oriented data structures. Each class object contains key words field and procedure for implementation of relationships based on their codes. The data models can include two or more related class objects. The process of information extraction, analysis and finally decision making through object oriented modelling has sequential character in spite of the hierarchical subjects’ classification of the model (figure 4).

The presented model aims at analyzing similar properties of the objects from definite class objects as well as differences toward treatment of crops against one or more pests. The analysis of these variations supports decision making. The logical sequence “crops–pests–pesticides” is implemented.

Let’s define: *A:* subclass of class “Crops”; *B:* subclass of class “Pests”; *C*: class objects “Pesticides”. The data modelling can be presented as follows:

a) Input data: Choose N object instances of *A;* Choose M object instances of *B*.

b) Apply operations: $(A \longrightarrow B) \longrightarrow C$. These two operations indicate a code application for implementing relationships between instances of the defined subclasses *A, B* and class *C*. For each instance of *A* and each instance of *B* will be found corresponding subset of *C*. The result is a matrix with NxM sets:

$$D = \begin{vmatrix} D_{11} & D_{12} \dots D_{1M} \\ \dots\dots\dots\dots\dots \\ D_{N1} & D_{N2} \dots D_{NM} \end{vmatrix}$$

The set D_{ij} (i=1, 2, … N; j=1, 2, … M) contains pesticides which are proper for use in treatment i^{-th} crop against j^{-th} pest. All sets of the matrix *D* are subsets of the set *C*.

c) Section (set *E*) and disjunction (set *F)* of the chosen sets of the matrix *D*. Set *E* shows the common pesticides for treatment of the chosen crops and pests. Set F refers to all pesticides.

This model can be applied for each subclass of objects and instances belonging to classes "Crops" and "Pests". Let's present an example. Two crops {"cucumber", "tomato"} are object instances of class objects "Vegetables" while pest {"mildew"} is instance of class objects "Fungi". Table 1 presents the final result of data processing – common pesticides for treatment the both crops {"cucumber", "tomato"} against {"mildew"}.

Table 1 Common pesticides for treatment "Cucumber" and "Tomato" against "mildew" (set *E*)

№	Code	Pesticides_Company	Active Ingredient	Dose
1	23	Bravo 500 Syngenta	500 g/l chlorothalonil	0.3%
2	51	Equation DuPont	225 g/kg famoxadone + 300 g/kg simoxanile	0.04% (40 g/da)
…	…	…	…	…
8	196	Champion Nufarm	77% cupric hydroxide	0.15%

The modelling based on reverse relationships "pesticides–pests–crops" is useful because of extraction information on available pesticides and variances for their practical use. The object oriented approach provides extending the information model (figure 4), e.g. to add new class objects in hierarchical structures. The model extension can be done in the sense of cognitive and application aims of the IS including biological, geographical and other factors. The flexibility of the model allows the development of application data models focused on different users groups:

- Agronomists and farmers responsible for crop protection measures. Instance level modelling allows working with real data on cultivated crops. The agronomists need to know the variations for the use of different pesticides against identified pests;
- Scientists and experts in the fields of biology, chemistry and phytopharmacy. They would be provoked by the instance level modelling and real data analysis in two directions: research and development of more effective chemicals for crop protection and improvement of the governmental rules, regulations and control mechanisms;
- Specialists who manage pesticides business and related technical and financial resources. The expert analytical information for needed pesticides is important for applying economical models: managing inventories, cost benefit and risk assessment.

The analogy between classes and subclasses in biology and chemistry and class objects and instances of object oriented approach is a base for flexible data modelling. The modern trends and problems coming from the field of crop protection require extending of the models through adding data and new class objects referring to the variations of pest resistance to a pesticide [13], changes in the nature (soil and climatic) etc. The object oriented models can be developed not only on the base of new classes of chemicals, but also by classes of natural products for crop protection.

4 Conclusion

This paper presents a conception for building and exploring of IS based on related text data through integration of two basic approaches – relational for data structuring and object oriented for data analysis. This conception is successfully applied for the creation of Bulgarian information system in the field of crop protection. The potential of the object oriented modelling consists of creating agile models based on related class objects. The developed solution provides working with easy retrieval database, flexible data models and object instances for data analysis. The extracted information will enhance expertise of the specialists and will facilitate decision-making process. Observing specific regulations referring to the pesticides use strategy, it can be concluded that data in the field of crop protection are different in different countries, but the relational database structuring and the software for object oriented modeling will be very similar. In this sense the presented ideas and experience can be useful.

References

1. Bulgarian Ministry of Agriculture: List of the permitted products for plant protection and fertilizers in Bulgaria, Sofia, Videnov & son (2009).
2. Feldman, R., Sanger J.: The text mining handbook. Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press (2007).
3. Bramer M.: Principles of data mining, Springer-Verlag London Limited (2007).
4. Crowsey, M., Ramstad, A., Gutierrez, D., Paladino, G., White, K.P.: An Evaluation of Unstructured Text Mining Software. In: IEEE Systems and Information Engineering Design Symposium, Charlottesville (2007).
5. Mahgoub, H., Rösner, D., Ismail, N., Torkey F.: A Text Mining Technique Using Association Rules Extraction, International J. of Computational Intelligence, Vol. 4, pp. 21-27 (2007).
6. Mooney, R., Bunescu, R.: Mining Knowledge from Text Using Information Extraction, Natural language processing and text mining, ACM SIGKDD Explorations Newsletter, Vol. 7, pp. 3 –10 (2005).
7. Lopresti, D.: Optical character recognition errors and their effects on natural language processing. In: Second workshop on Analytics for noisy unstructured text data, ACM Digital Library, 303: pp. 9-16 (2008).
8. Onkov, K.: Effect of OCR-errors on the transformation of semi-structured text data into relational database. In: Third Workshop on Analytics for Noisy Unstructured Text Data, pp. 123-124, ACM Press, New York, (2009).
9. Dimova, D., Onkov, K.: An algorithm for automated creation of a PC database storing related text objects, Journal of Information Technologies and Control, 5(2), 48-52 (2007).
10. Embley, D.W., Campbell, D.M., Smith, R.D., Liddle, S.W.: Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. In: Seventh International Conference on Information and Knowledge Management, pp. 52-59 (1998).
11. Necasky, M.: Conceptual Modeling for XML, IOS Press (2008).
12. Kouno, T., Ayabe, M., Hitomi, H., Machida, T., Moriizumi, S.: Development of Relational System between Plant Pathology Database and Pesticide Database, In: Second Asian Conference for Information Technology in Agriculture, AFITA (2000).
13. United States Environmental Protection Agency: http://www.epa.gov/
14. Bulgarian Ministry of agriculture, Authorized institution for plant protection, www.stenli.net/nsrz/main.php?module=info&object=info&action=view&inf_id=21