Serdica J. Computing **4** (2010), 487–504

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

AN ALGORITHMIC SOLUTION FOR MANAGEMENT OF RELATED TEXT OBJECTS WITH APPLICATION IN PHYTOPHARMACY*

Delyana Dimova†

Abstract. This paper presents an algorithmic solution for management of related text objects, in which are integrated algorithms for their extraction from paper or electronic format, for their storage and processing in a relational database. The developed algorithms for data extraction and data analysis enable one to find specific features and relations between the text objects from the database. The algorithmic solution is applied to data from the field of phytopharmacy in Bulgaria. It can be used as a tool and methodology for other subject areas where there are complex relationships between text objects.

ACM Computing Classification System (1998): D.0, H.2.4.

Key words: Algorithmic solution, relational database, data processing, software, phytopharmacy.

*This article presents the principal results of the doctoral thesis "An Algorithmic Solution for Management of Related Text Objects with Application in Phytopharmacy" by Delyana D. Dimova, defended before the Specialized Academic Council for Electronic and Computing Technologies and approved by the Supreme Attestation Commission, 2008.

†I would like to express my gratitude to my research adviser Assoc. Prof. Dr K. Onkov and the reviewers of my dissertation Sen. Researcher Dr A. Eskenazi and Prof. Dr N. Kolev.

Introduction. Automatic text analysis may involve different levels of information embedded in the text. This motivates many subtasks and modules. Since the topics covered in a document are usually related to the context of the document, analysing topical themes within this context can potentially reveal many interesting theme patterns [1]. It should be noted that documents used in practice are often not submitted in electronic form and their processing is largely hampered, requiring the use of OCR systems [2, 3]. Optical Character Recognition works by scanning source documents and performing character analysis on the resulting images, producing a conversion to encoded text, which can then be stored and manipulated electronically like any standard electronic document. The importance of retrieving OCR text efficiently has grown significantly in recent years [4].

The text-level structure of each document is described by the occurrence of typical functional components in the text [5]. The problem of text mining, i.e., discovering useful knowledge from unstructured text, is becoming an increasingly important aspect of KDD (Knowledge Discovery from Databases) [6]. KDD considers the application of statistical and machine-learning methods for discovering novel relationships in large relational databases [7]. Unfortunately, for many applications, available electronic information is in the form of unstructured natural-language documents rather than structured databases [6]. Many text documents used by specialists and experts contain text objects (words, phrases, sentences, etc.) between which there are complex relations. In the process of search and extraction of all relations between the text objects of the documents several difficulties arise. In text mining [8, 9] or information retrieval systems visualization methods can improve and simplify the discovery or extraction of relevant patterns or information. Many of the approaches developed for text mining purposes are motivated by methods that had been proposed in the areas of explorative data analysis, information visualization and visual data mining [10, 11]. Data mining refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in relational databases. Of particular importance is the development of powerful tools [12] for analysis and interpretation of these data and to extract knowledge which helps users in solving complex problems. IE systems can be used to directly extract abstract knowledge from a text corpus, or concrete data from a set of documents which can then be further analysed with traditional data-mining techniques to discover more general patterns [7].

This work studies problems of such nature and provides an integrated solution for management of text objects (words, phrases) and relations between them.

The aims of the paper are:

– To present an algorithmic solution for management of related text objects which is a powerful tool for their extraction from documents in electronic or paper form, for their storing, structuring and processing in a relational database.

– To present an application of the developed algorithmic solution and applied software for the data in the field of phytopharmacy in Bulgaria.

Algorithmic solution. The algorithmic solution for management of related text objects includes (Figure 1):

A) Optical text recognition. Despite the good quality of the FineReader software, errors occur during its use. Therefore an algorithm for automated error correction after optical recognition by FineReader [13] was developed. The purpose of this algorithm is to reduce the number of errors from optical character recognition and control of input data.

B) A relational model for presenting related text objects (words, phrases).

C) – A system for coding related text objects [14] that is used for coding words and phrases in electronic documents, as well as for designation of relationships between them.

– An algorithm for automated creation of a relational database [14]. Text objects from the electronic documents are extracted, coded and entered into a relational database through this algorithm. The algorithm developed has the properties of an IR (information retrieval) approach used in text mining.

D) Algorithms and applied software for data extraction and data analysis from the relational database. Applied software is built for the purpose of finding, comparing, analysing and summarizing the special features and relations between the related text objects. In these cases some typical approaches of data mining are applied.

Algorithms, a relational model and database design. *Algorithms, building the developed solution:* This work studies documents that contain text objects between which complex relations exist are studied. M_1 and M_2 are sets that contain these text objects. Each text object is a word or phrase (sequence of words). It is possible that one of the sets consists of subsets which do not intersect. Let $M_2 = P_{M21} \cup P_{M22} \cup P_{M23}$ and $P_{M21} \cap P_{M22} \cap P_{M23} = \emptyset$. Each object of a set may be related with one or more objects of the other set. The text objects of the sets M_1 and M_2 should be extracted from the documents and stored in a relational database.

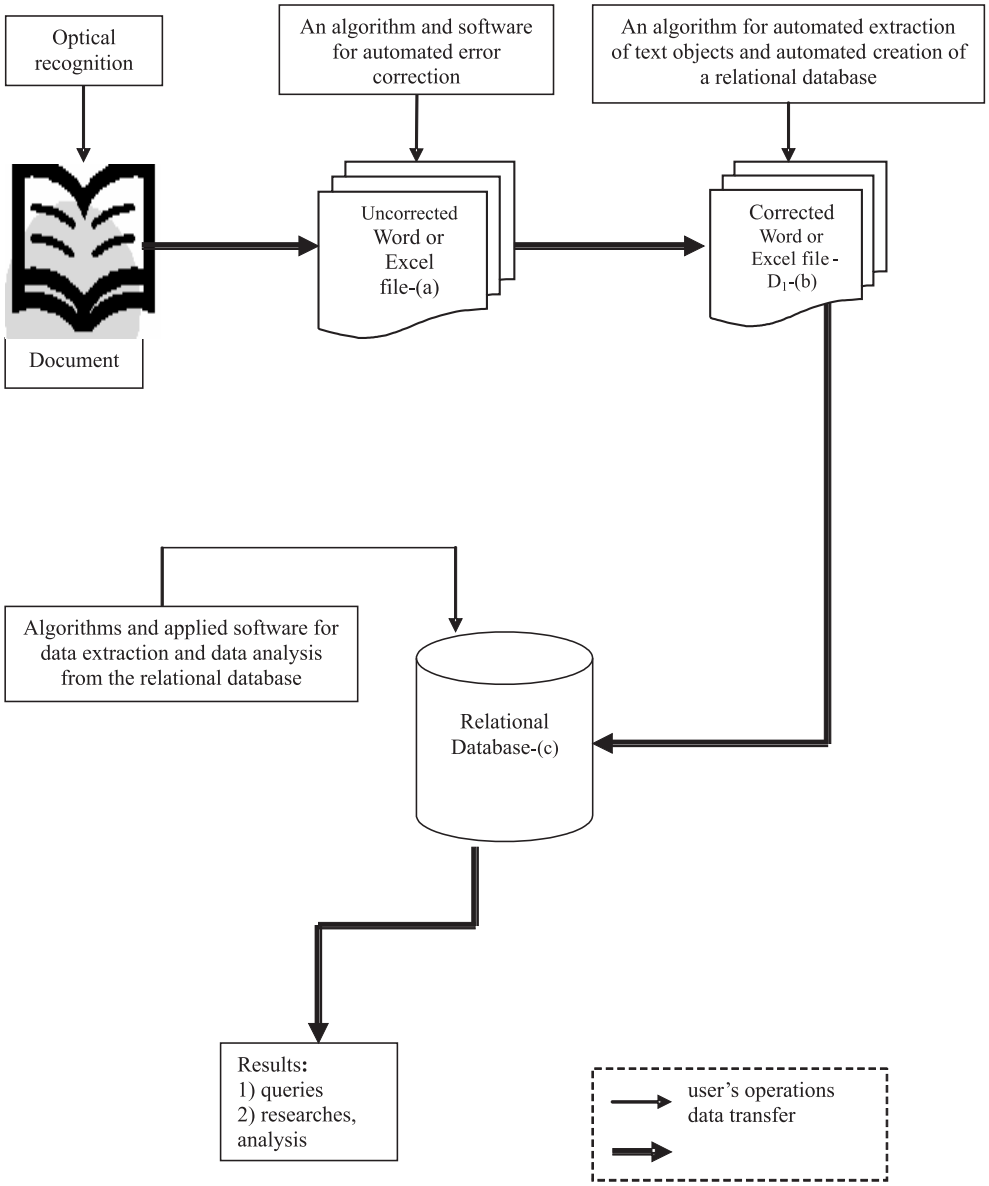


Fig. 1. Scheme of the developed algorithmic solution for management of related text objects

The developed algorithmic solution includes OCR software and an algorithm for automated error correction after optical recognition by FineReader [13]. The FineReader system is used for documents that contain related text objects, but are presented in paper form (Figure 1–a). In the process of their transforming into electronic form errors occur after the optical recognition. The typical errors in optical character recognition can be classified into two groups.

– Incorrect recognition of various characters - numbers written in Roman and Arabic numerals, hyphenated words, units of measurement (cm, m, da, etc.), punctuation marks;

– Discrepancy of data arranged in tabular form.

Particular attention should be paid to errors (for example hyphenated words) in key words. If they are not corrected and are objects belonging to the M_1 or M_2, then they will not be found in the electronic document and input in the relational database. The output file obtained as a result of the corrections can be in Word or Excel format (Fig. 1–a). All errors of the optical character recognition may not be corrected, but the developed algorithm for automated error correction significantly reduces their number.

When building the relational database a coding system of the text objects [14] is used for the following reasons:

– Working with text objects (words, phrases) particularly within a large volume of documents is extremely cumbersome and difficult. This entails replacing them with natural numbers (codes). For each object a relevant code is introduced. Moreover, natural numbers are indexed more easily than text. As a result a faster and more efficient search of text objects is performed.

– Some codes are used as keys in a relational database.

– Codes are used to designate the relations between objects.

Let D_1 be an electronic document (Figure 1–b), which contains the text objects of the sets M_1 and M_2 described above. An algorithm was developed for the extraction of these objects and their storing in the relational database [14]. Exactly one natural number (code) corresponds to each text object. The codes of the objects are introduced into the lists in advance with the objects themselves. A text object in a row of D_1 can be recognized as an element of the list containing text objects from M_1 or M_2. Then its code is stored in the row of the document. Relationships between objects in the row of the document are identified by the formation of ordered pairs of the type (k_1, k_2), where k_1 is a

code of the object from M_1 and k_2 is a code of the object from M_2. This is why all codes of objects from both sets are combined. The position of the elements in the ordered pair does not indicate the direction of the relationship between the corresponding text objects. The direction of the relationship of related objects is determined depending on the particular subject area database. Data that have been found in the electronic document were entered in the corresponding tables of the relational database (Figure 1–c). The algorithm for the automated creation of the database can also be used for updating if new objects from an electronic document have to be added to the database.

The algorithmic solution contains algorithms and software for processing the relational database. An algorithm was developed for extracting the related text objects from the sets $_1$ and $_2$, which are presented in various tables of the database. In this case for a selected object from one set all combinations of its code and codes of objects of the other set are searched. Depending on the codes of objects of the other set that are found, their corresponding objects are determined. As a result, all related text objects from both sets can be extracted. In practice, this leads to finding all records: $h \subset M_1 \times M_2 = \{(a_1, b_2)/a_1 \in M_1, b_2 \in M_2)\}$ or $g \subset M_2 \times M_1 = \{(c_2, v_1)/c_2 \in M_2, v_1 \in M_1)\}$. The algorithms for data extraction and data analysis contain subalgorithms, which implement some of the most frequently used relational algebra operations—join, projection and selection.

As a result of the implementation of these operations subsets of columns or rows of a table can be extracted, data from multiple tables can be brought together in one table, and specific dependencies and relations for the text objects (words, phrases) can be found as well.

The effectiveness of the developed software can be shown in the application to the related text objects in the field of phytopharmacy that are discussed in the next section.

A relational model of the related text objects: a database design. The extracted text objects from the electronic document D_1 are stored in tables of the relational database. When designing the relational database the number of tables will depend on:

- The set M_1.
- The number of the subsets of the set M_2 (in the case – P_{M21}, P_{M22}, P_{M23}).

Then the number of tables will be $1 + q$ (with q denoting the number of subsets of M_2). The table T_M stores the text objects of the set M_1. The defined

primary key in T_M coincides with the codes of the objects of M_1. The text objects of M_2 are stored in separate tables (T_{P1}, T_{P2}, T_{P3}). They are related to the objects of the table T_M through the codes of the objects of M_1. The set of codes of the objects of M_1 is a primary key in the referenced table T_M and a foreign key in the referencing table T_{P1}, T_{P2} or T_{P3}. All relationships between the tables in the relational database are of "one to many" type. The requirement that each value existing for the foreign key be a corresponding value in the referenced key must be fulfilled for the designed relational database. The three related tables consisting of text objects of the subsets P_{M21}, P_{M22} and P_{M23} also contain all combinations of the codes respectively of an object from M_1 and the object from the subset. This means that the code of the text object from the set M_1 exists. It has been previously entered into the referenced table. Its value in the related table is the value of the foreign key. Thus the referential integrity is guaranteed. The rule which refers to the entity integrity is also fulfilled. The attributes that are primary keys cannot take null values. In the tables of the relational database this key belongs to the set of natural numbers. The degree of the tables T_M, T_{P1}, T_{P2}, T_{P3} cannot be determined in advance. They may contain other attributes for text objects depending on the specific subject area database created.

There are cases in which the values of the number q may be much larger than the one discussed above ($q = 3$) and the text objects of the set M_2 can be related with text objects of another set. It is therefore expedient for the objects of each set to be presented in only one table. Otherwise the number of tables in the relational database can become quite large. Therefore the scheme of the relational database can be much more complicated and its efficiency would be reduced. This makes it difficult to use the relational database.

It is possible for the same relationships, which are defined between the text objects of M_1 and M_2, to also be defined on the text objects from M_2 and the other set M_3, etc. Then the text objects from M_3 should be introduced in the database. For this purpose the coding system is used. The difference here is that in addition to relations between M_1 and M_2 there are also relations between M_2 and M_3. These relationships between the text objects must be found in the input document. For this purpose two ordered pairs (k_1, k_2) and (k_2, k_3) are defined, where k_1, k_2 and k_3 are respectively the codes of text objects from the sets M_1, M_2 and M_3. The cases in which the sets are more than two, additional restrictions have to be introduced. The formation of an ordered pair of codes of objects of different sets between which there are no defined relationships should not be allowed.

Application of the algorithmic solution and applied software for the data in the field of phytopharmacy in Bulgaria. Phytopharmaceutical products (pesticides) are made publicly known in the reference book *List of authorized and employed plant protection products on the market* [15] in Bulgaria. The pesticides are divided into three groups—fungicides, insecticides and herbicides. They are arranged alphabetically in the reference book. Their active substances, which are listed separately, are arranged in the same way. The columns contain data for the characteristics of the pesticide. For each pesticide are given: the name of the pesticide and the company name maker; active substance; concentration for use; pest/culture where it is applied; average lethal dose; quarantine period in days and category of use (Figure 2)[1]. The use of the reference book raises problems concerning:

- The search for all relationships between text objects (culture, pests or/and pesticides). These relations are presented in the publication, but are found on different pages, requiring much time and effort.
- The possibility of missing a certain object or other objects related to this object.
- Difficulties in decision making and giving a quick and correct answer to a question.

	Pesticide, company	Active substance	Concentration	Culture/pest against which the pesticide is used.	LD	Quarantine period in days	Category of use
18	BENIMOST 50 VP Hokley	50% benomil	0,1 %	gray mold on grapes; powdery mildew on peach	5000	28	1
19	BENOMIL 50 V Sinon	500g/kg benomil	0,1 %	powdery mildew on peach; gray mold on grapes; powdery mildew on tomatoes	5000	28	1

Fig. 2. Presentation of the data in the reference book

[1]In the reference book the original text in Figure 2 is in Bulgarian.

The shortcomings mentioned in the use of the reference book of permitted phytopharmaceutical products could be avoided by applying the developed algorithmic solution for the management of text objects and a software system. A compact presentation of the pesticides, cultures and pests in a relational database and the application of software for their extraction and analysis that are included in this system would facilitate the user's experience significantly. This would lead to the faster solving of problems encountered by specialists and experts in the field of the agriculture (plant protection, agronomy and others). The software we need for applying the developed algorithmic solution is implemented in the environment of Visual Basic.

Practical realization of the model: The reference book is used in building the "Phytopharmacy" database. Data for the permitted pesticides are interpreted as follows (Figure 3):

- The set M_1 contains all cultures. It is denoted as the set "Plants";
- The set M_2 ($M_2 = P_{M21} \cup P_{M22} \cup P_{M23}$) consists of all pests (fungal diseases—fungi, weeds and insects). It is denoted as the set "Pests":
  - $P_{M21} \equiv$ "Fungi". The subset "Fungi" contains all fungal diseases.
  - $P_{M22} \equiv$ "Weed". The subset "Weed" contains all weeds.
  - $P_{M23} \equiv$ "Insect". The subset "Insect" contains all insects.
- The set M_3 that was presented in the description of the relational model of related text objects consists of all permitted phytopharmaceutical products. It is denoted as the set "Pesticides".

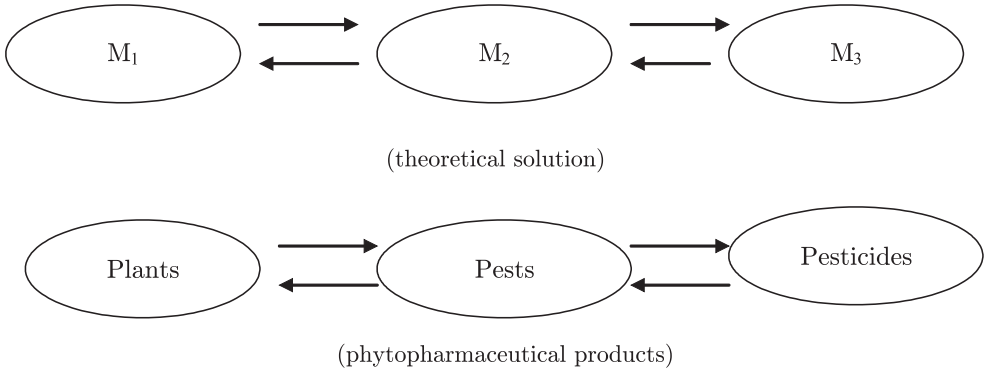


Fig. 3. Sets containing the text objects

The reference book of permitted pesticides is transformed into electronic form, in Excel files, by using ABBYY FineReader. The structure of the xls file is similar to the one presented in Figure 2. The correction of the typical errors discussed above which occurred after the optical character recognition is done by the algorithm and software for automated error correction. Pests and cultures in the described sets and their codes are introduced beforehand in lists. Exactly one code (natural number) corresponds to each culture or pest.

The software for automated extraction of objects from the Excel files and automated creation of the database performs the following operations:

– For each culture or pest in a row of the Excel table, their codes from the pre-selected list are introduced.

– All possible combinations of the codes of objects belonging respectively to the sets "Plants" and "Pests" are formed. Each row of the Excel file contains a pesticide and the related pests of a culture or a pest of several cultures. In this case the combinations in the row of the Excel file are also applied to the designation of the relationships between the pesticide and the two objects (culture and pest) found.

– All the objects found from the sets "Plants", "Pests" and "Pesticides" are extracted and entered into the relational database "Phytopharmacy".

Design of the Relational Database "Phytopharmacy": Data for the phytopharmaceutical products are systematized in separate tables. The "Phytopharmacy" database [16] contains the following relational objects (Figure 4):

- The table "Agr_Cultures"(id_C_ob, Culture_ob) stores all cultures of "Plants" and a primary key (the set of the codes of the cultures).
- The tables "Fungus_P" (Cod_F_id, Pest_F, Cod_ Pest_F, id_C_ob),
  "Insect_P" (Cod_I_id, Pest_I, Cod_ Pest_I, id_C_ob),
  "Weed_P" (Cod_W_id, Pest_W, Cod_ Pest_W, id_C_ob)
  contain the three types of pests, their codes and the codes of cultures to which they are related.
- The table "Pesticide"(id, Pesticide_firm, Active substance, Concentration, LD, Quarantine period, category, Features, Import, Note, Cod_F_id, Cod_I_id, Cod_P_id) stores all phytopharmaceutical products of the set "Pesticides" and their attributes.

The main difficulties in using the reference book for permitted pesticides were associated with a comparatively large number of complex relationships be-

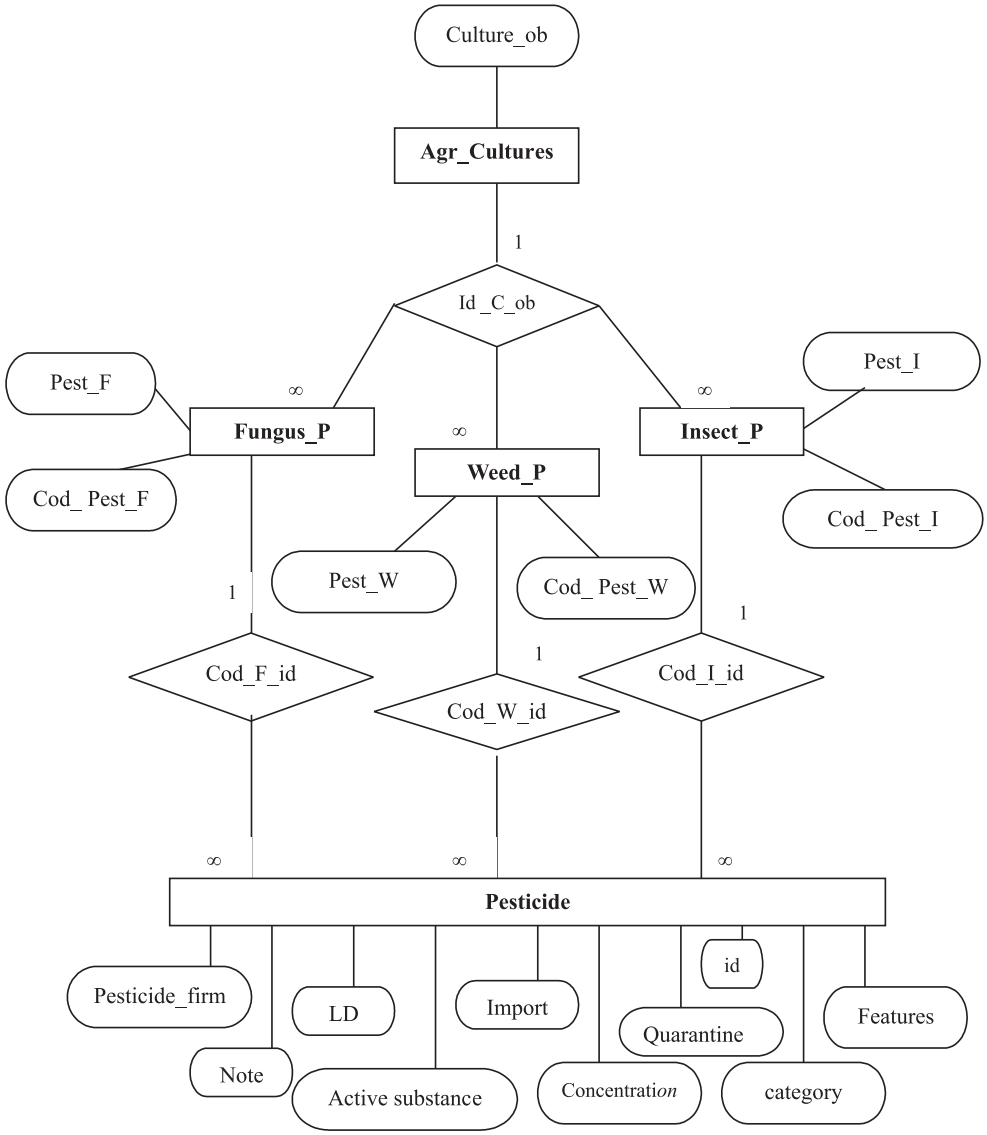


Fig. 4. The entity-relationship diagram of the "Phytopharmacy" database

tween the participating text objects (culture, pest, pesticide). The developed relational model of the text objects and the "Phytopharmacy" database solve the problems concerning storing, access, data processing and ensuring entity integrity. The reference book for permitted pesticides is reissued every year. The main reason for this is the constantly changing requirements for the phytopharmaceutical products. These changes must be reflected in the objects of the database. Therefore an annual update should be performed. The DBMS "Access" is used as a platform for the development of the "Phytopharmacy" database.

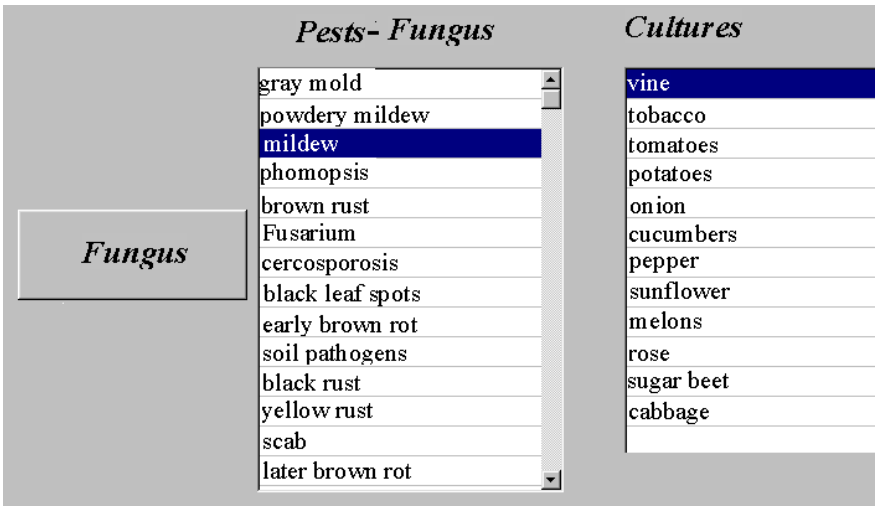
Data extraction from the relational database "Phytopharmacy": The applied software is developed for the purpose of quick and easy processing of the text objects stored in the relational database "Phytopharmacy". It is primarily intended for use by agronomists, specialists in plant protection, but also by farmers and others. This software allows users to extract the necessary data and relations between objects from the database in the following order [17].

- Lists consisting of the names of cultures, pests (insects, weeds or fungi) or phytopharmaceutical products with their characteristics (active substance, lethal dose, category of use, etc.)
- Presentation of data for two related text object. Data referring to the permitted phytopharmaceutical products contain the following combinations: "cultures–pests", "pests–pesticides", "pesticides–pests" and "pest–cultures" (Figure 5–a);
- Data on three related text objects. In this case the relations are "cultures–pests–pesticides" and "pesticides–pests–cultures" (Figure 5-b)[2].

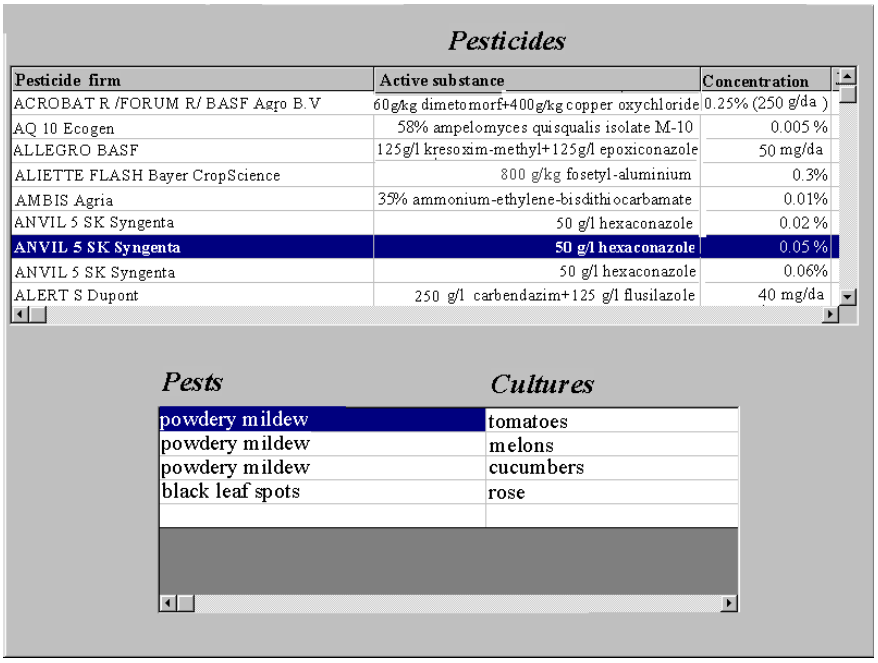
In addition, for each object (culture, pest, pesticide) the number of objects from the relational database related to it can be found by the developed software. This information is intended mainly for use by a significantly narrower range of specialists in the field of agriculture who are principally engaged in the study of diseases in cultures.

Data analysis of pesticide, culture, pest and possible alternatives: Dedicated software tools were developed, especially for data analysis of the relational database "Phytopharmacy". Thes software tools are designed to search for specific features of phytopharmaceutical products, mostly relationships with cultures and pests, in order to study possible alternatives in decision-making.

[2]The data about the phytopharmaceutical products shown in figure 5 are in Bulgarian in the original text.



a) Data referring to the following combination: "pest–cultures"



b) Data for "pesticide-pests–cultures"

Fig. 5. Presentation of the data for the phytopharmaceutical products

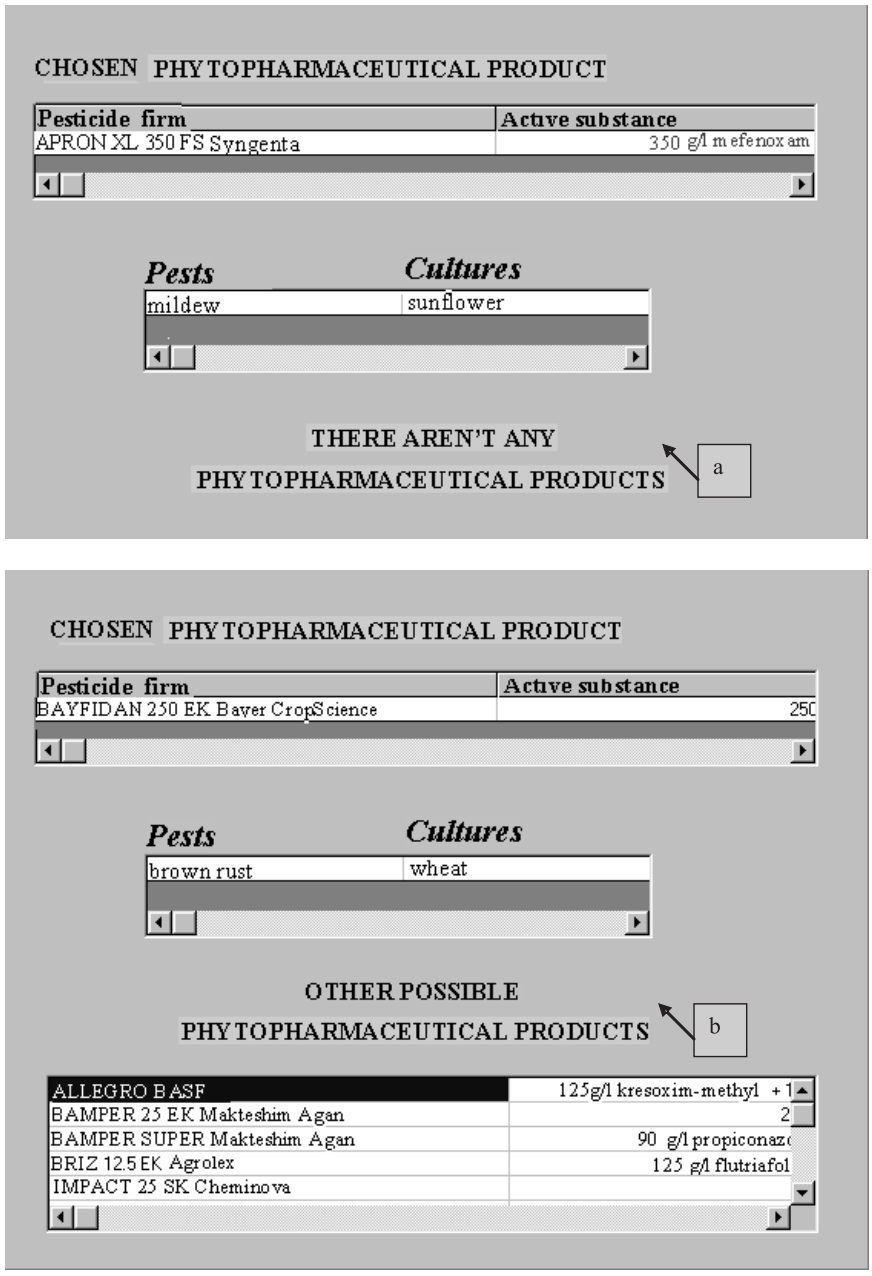


Fig. 6. Data analysis for the chosen pesticide

When choosing a pesticide related to cultures and pests from the database the following alternatives are suggested:

- There is no other phytopharmaceutical product for the treatment of the culture against the pest except the presented one. As a result, there is no chance of replacing it by other pesticides (Figure 6–a). Regardless of its effectiveness when it is used against the pest, the users practically have no other choice.
- There are other pesticides presented which could replace the pesticide chosen to be applied against the pests of the culture (Figure 6–b)[3]. Then it is possible to perform the following types of research:
  - Concerning the existence of values of the dose (concentration) different from the found value (respectively lower or higher), in which the chosen pesticide is successfully used for this pest in the culture. In these cases the pesticide can be combined with another. It should be noted that this condition is mentioned in the reference book in the relational database "Phytopharmacy".
  - For other phytopharmaceutical products used for the same culture and pest, the values for concentration, lethal dose, quarantine period in days and active substances can be compared. As a result, it is possible to carry out a replacement of the selected pesticide and presented opportunities. This choice could depend on the efficacy of the pesticides and numerous other factors.

The presented investigations of data of the phytopharmaceutical products are important for specialists in the agricultural field. The results obtained from the data analysis enable decisions concerning the use of pesticides with lower toxicity. In this way the negative impact of the products used is reduced, both for the environment and for humans and animals. It should be noted that in data analysis for pesticides the complex relations between them are studied primarily. The alternatives presented can be used for finding rational and adequate solutions in various problems of plant protection.

Conclusion. The developed algorithmic solution is a powerful tool for the management of the related text objects. The basis of the solution are integrated algorithms for automated extraction of the related objects from text

[3]The data about the pesticides, shown in figure 6 is in Bulgarian in the original text.

documents and their presenting in a relational database, as well as algorithms for data analysis and data extraction from the database. Elements of set theory are used in building the relational model of the text objects. The algorithms for data analysis and data extraction have important practical significance as well methods of relational algebra and coding system are applied in themselves. The algorithmic solution for the management of text objects has been researched and used for data (cultures, pests, pesticides) from the field of phytopharmacy in Bulgaria. Data from the reference book for permitted pesticides and cultures and pests related to them are presented in the relational database "Phytopharmacy". As a result the problems of searching, access and data processing referring to these related objects are solved. The developed software system includes the automated creation of the relational database "Phytopharmacy" as well as application software for data extraction and analysis of the database. This allows for the research of the pesticides and finding useful information for specialists in the field of agriculture, farmers and others. The algorithmic solution and the developed software can be applied to text documents containing related objects in areas such as pharmacy, medicine, etc.

REFERENCES

[1] Mei Q., C Zhai. A Mixture Model for Contextual Text Mining. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, Pennsylvania, USA, August 20–23, 2006, 649–655.

[2] `http://www.abbyy.com`

[3] Scanning text – user's guide Optical Character Recognition with OmniPage Pro. Scanning Text User Guide Aug05.doc, ©Arizona Board of Regents, 8.11.2005.
`www.library.arizona.edu/documents/ust/ScanTextGuideAug05.pdf`

[4] Beitzel S., E. Jensen, D. Grossman. A Survey of Retrieval Strategies for OCR Text Collections. Symposium on Document Image Understanding Technologies, Greenbelt, Maryland, April 2003.

[5] Kando N. Text structure analysis as a tool to make retrieved document usable. In: Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages, Taipei, Taiwan, Nov. 11–12, 1999, 126–135.

[6] Mooney R., R. Bunescu. Mining Knowledge from Text Using Information Extraction. Natural language processing and text mining. *ACM SIGKDD Explorations Newsletter*, **7** (2005), No 1, 3–10.

[7] Mooney J., U. Nahm. Text Mining with Information Extraction, Multilingualism and Electronic Language Management. In: Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa (Eds W. Daelemans, T. du Plessis, C. Snyman, L. Teck) Van Schaik Pub., South Africa, 2005, 141–160.

[8] Hearst M. Untangling Text Data Mining. In: Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20–26, 1999, 3–10.
`http://www.sims.berkeley.edu/~hearst`

[9] Hearst M. What Is Text Mining.
`http://www.sims.berkeley.edu/~heast/text-mining.html`, 2003

[10] Hotho A., A. Nürnberger, G. Paass. A Brief Survey of Text Mining, Zeitschrift fuer Computerlinguistik und Sprachtechnologie. *GLDRV – Journal for Computational Linguistics and Language Technologie*, 2005, 19–62.

[11] Don A., E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, C. Plaisant. Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In: Proceedings of the Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisbon, Portugal, 213–222.

[12] `http://www.exinfm.com/pdffiles/intro_dm.pdf`

[13] Dimova D., K. Onkov. Algorithmic solution for errors correction after optical recognition by FineReader. In: Proceedings of the International Conference Automatics and Informatics'04, Sofia, Bulgaria, 6–8 October 2004, 251–253.

[14] Dimova D., K. Onkov. An Algorithm for Automated Creation of a PC Database Storing Related Text Objects. *Information Technologies and Control*, **2** (2007), 48–52.

[15] Списък на разрешените за предлагане на пазара и употреба продукти за растителна защита. Виденов&Син ЕООД, 2007.

[16] Dimova D. Data model of application of the permitted pesticides in Bulgaria. In: Proceedings of the International Conference Automatics and Informatics'07, Sofia, Bulgaria, 3–6 October 2007, 15–18.

[17] Onkov K., D. Dimova. Information Technology for Creation and Use of the PC "Phytopharmacy" Database. In: Proceedings of the International Conference "Information Systems in Sustainable Agriculture, Agroenvironment and Food Technology", Volos, Greece, 2006, 10–16.

Delyana Dimova
Department "Computer science and Statistics"
Agricultural University
12, Mendeleev Blvd.
Plovdiv, Bulgaria
e-mail: delyanadimova@abv.bg

Received April 26, 2010
Final Accepted July 26, 2010