KNOWLEDGE ACQUISITION FROM AGRICULTURAL DATABASES*

J. Vaníček, V. Vostrovský

Czech University of Life Sciences, Faculty of Economics and Management, Department of Information Engineering, Prague, Czech Republic

Knowledge and information are the key to successful pursuit of business activity. For the management knowledge can be employ knowledge acquisition from databases. This paper describes possibilities of the knowledge acquisition from the agricultural databases by means of the association rules generalization. These features are showed at the example of the herbicide list. This paper denotes the usability of the association rules generalization method in our agriculture. In this manner knowledge acquisition can be valued support for the effective buy of the protective means.

knowledge; expert system; relational database system; association rule; knowledge acquisition

INTRODUCTION

Recently, the problem of the knowledge acquisition and efficient knowledge exploitation is very popular also in agriculture area. Various methods of knowledge acquisition are examined.. In this context the database methods are in the centre of interest. The aim of this activity is the knowledge extraction and acquisition from the existing databases. The main aim of this article is to demonstrate the possibilities of knowledge acquisition from agriculture databases by means of association rules. These possibilities will be illustrated on the case of the methodical manual for plant protection published by the Ministry of Agriculture of the Czech Republic.

MATERIAL AND METHODS

Knowledge acquisition from database records can be characterized as a not trivial acquisition of earlier unknown and potentially useful information in the implicit form, from these records. (F a y y a d , I r a n i , 1993). The source can be some existing relational databases, consisting of the set of relations, representing two-dimensional tables, with rows corresponding to recording entities and columns corresponding to attributes of recording entities. The entities recorded as the rows in relations can be often interpreted as simple rules represented a corresponded knowledge. An example from the agriculture area can be the methodical manual for plant protection published in the printed form by the Ministry of Agriculture of the Czech Republic. The records containing in the mentioned manual can be simple transformed into the corresponding relation tables of the form presented in Table 1.

One of the methods for knowledge acquisition from the existing databases is the method of association rules generalization. In the presented article this method is applied for the priority setting for various producers of plant protective preparations in the case of concrete weed occurrences for Czech farmers. The association rules has been studded for instance by A g r a w a l et al. (1996) for consumer's basket analysis.

The necessary condition of such an acquisition of other knowledge from existing databases is data preprocessing of records from these data sources. The aim of this data pre-processing is:

- the selection relevant data for the knowledge acquisition,
- the representation of these data in the suitable form for the processing using the specific algorithm (B e r k a , 2003).

For the pre-processing purpose the relational database technology provides the relative efficient tool – the query language SQL for the definition of tables and manipulation with data. The SELECT statement for the selection and CREATE VIEW for the virtual relations creating using the join of several existing relations can be used for this purpose. The possibilities of this pre-processing can be dem-

Table 1. Relational table containing the set of registered preparations for plant protection

Protective preparation	Harmful factor	Growth	Dose kg/ha	Protection period	Application day
TOLKAN FLO	couch grass	sugar-beet	1.5	AT	postharvest
MARATON	hair-grass	rye	0.015	AT	during autumn
ARELON 500	catchweed	spring barley	0.25		during spring
REFINE	redshank	meadowland	0.015		during spring

* Supported by the Ministry of Education, Youth and Sports of the Czech Republic (Grant No. MSM 6046070904 – Information and knowledge support of strategic control and No. 2C06004 – Information and knowledge management – IZMAN).

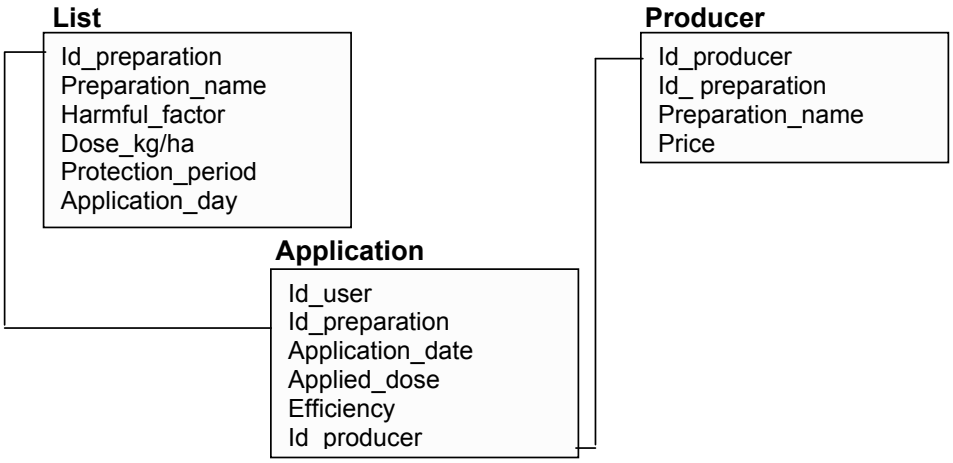

Fig. 1. Database logical schema of registered protective preparations and their applications

onstrated for instance on the case of the register of preparations for plant protection published in the above mentioned manual by means of the following operations (Fig. 1):

```
SELECT id_preparation, preparation_name, harmful_
        factor, applied_dose, application_date,
        efficiency FROM list, application, producer
        WHERE list. id_preparation =application. id_
                preparation
                AND application.id_
                producer=producer.id_producer;

CREATE VIEW application_efficiency AS SELECT Id_
        preparation, preparation_name, harmful_
        factor, applied_dose, application_date,
        efficiency
        FROM list, application, producer
                WHERE list. id_preparation =
                application. id_preparation AND
                application.id_producer = producer.id_
                producer AND efficiency > 50;

SELECT * FROM application_efficiency;
```

The more broad pre-processing possibilities offer the programming enhancement of the SQL language, the language PL/SQL, which allows realizing the simple variant of the inference mechanism used in the classical expert systems.

The list of the registered preparations mentioned above contains in the essence the set of production rules of the IT *E* THEN *H* types, in which *E* and *H* are two propositions in the implication. These propositions can be interpreted as evidences or hypothesis from the table. From the table 1 above can be for example derived the following rule

IF harmful factor (couch grass) $\wedge$ growth (sugar-beet)
THEN protective preparation (TOLKAN FLO) $\wedge$
dose kg/ha (1,5) $\wedge$ application day
(postharvest),

This rule can be interpreted as follows: For the couch grass eliminated from sugar-beet can be recommend protective preparation TOLKAN FLO dosed by 1.5 kg/ha with post harvest application.

The protective preparations, which can be selected, are components of the natural taxonomy. Individual producers offer the scale of these such preparations, with different level of effectiveness and different price (Fig. 2).

Such a taxonomy can be used for the generalized associating rules choosing (A g r a w a l et al., 1993), describing the associations between the entries on the different levels. All rules which can be derived from such evidences have the general form: *Ant* $\Rightarrow$ *Suc*, where *Ant* (the left part of the rule) is the assumption and *Suc* is the conclusion. For the grower the rules on the bottom level of hierarchy can be interesting. For example:

hair-grass in the wheat $\Rightarrow$ MARATON BAYER

However, also the rules on the more general level can be also more interesting. For example:

The knowledge then in the case of hair-grass occurrence the growers prefer the protection preparations from the firm BAYER.

The rules, generalized in such a way have the form *Ant* $\cap$ *Suc* = $\varnothing$ and no entry of *Suc* proceeds to any entry from *Ant* with respect to the existing hierarchy. The problem during the generalized rules finding consists in the choice of the minimal experimental support by the number of objects complained both the assumption and conclusion (B e r k a , 2003). If this support is exceedingly large, the rules on the bottom level are missing. If it is too low, we obtain an extensive and vast set of combinations.

If the rule *Ant* $\Rightarrow$ *Suc* satisfy the required support and required level of reliability then only for the rule *Ant* $\Rightarrow$ *predecessor*(*Suc*) there is guarantied that both two parameters satisfy this rule. There is any a priory evidence concerning the reliability of the rules *predecessor (Ant)* $\Rightarrow$ *(Suc), predecessor(Ant)* $\Rightarrow$ *predecessor*(*Suc*) (B e r k a , 2003). The basic association rule characteristics are the support and reliability, whereas the support is the number of objects satisfying the presumption and conclusion, and the reliability is the conditional probability of the conclusion, if the presumption is true (A g r a w a l et al., 1993). These features can be illustrated by means the example presented in the following table containing the realized

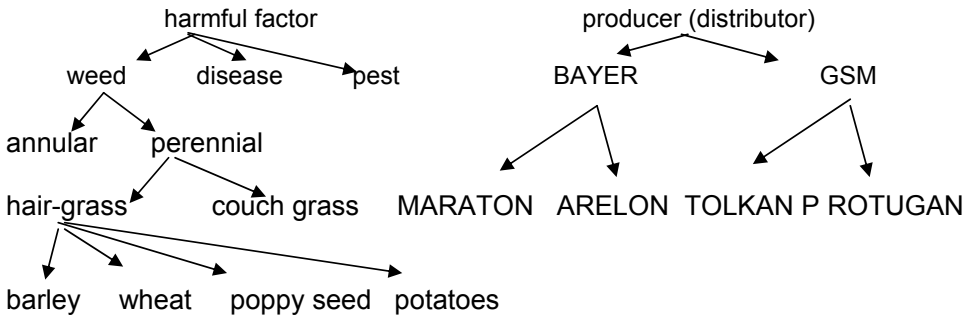

Fig. 2. Weed's taxonomy in plants and producers of the protective preparations

Table 2. Relation table with data applications of concrete protective preparations

Application day	Harmful factor – main growth	Protective preparation – producer
12. 3. 2005	hair-grass – wheat	MARATON – BAYER
12. 6. 2005	hair-grass – poppy seed	ARELON – BAYER
13. 5. 2005	couch grass – poppy seed	PROTUGAN – GSM
20. 6. 2006	hair-grass – barley	MARATON – BAYER
21. 6. 2006	hair-grass – potatoes	TOLKAN – GSM

application of the concrete protective preparations by the concrete growers.

From Table 2 can be determine the following entries counts:

hair-grass 4
couch grass 1
BAYER 3
GSM 2

From this evidence then can be deducted also the following generalized association rule

$$\text{hair-grass} \Rightarrow \text{BAYER}$$

with 60% level of support and 75% reliability. This rule can be explicated as follows: our growers prefer the protective preparations from firma BAYER an example occurrence of the hair-grass in grown.

RESULTS AND DISCUSSION

The main problem for association rules generalization is the question of the support choice on the different hierarchy level, which can be solve by means of required support changes depending on the level hierarchy flowing from know as commonness level g (Chung, Lui, 2000). This commonness level is defined by as proportion of the number of leaves on the subtree of the considered value and the number of all leaves in the hierarchy. Hence for example in the taxonomy mentioned on Fig. 2 the value *hair-grass* has the commonness level 4/7 and value BAYER has the commonness level 2/4. The commonness level *g(Comb)* of a combination is minimal commonness level of the single category in the combination (Berka, 2003).

The user define parameters minsup_0 and minsup_1, where minsup_0 is required support for the combination with the lowest commonness level (i.e. $g = 0$ and minsup_1 is required support for the combination with the commonness level equal 1 (i.e. $g = 1$). From these values the support of arbitrary combination as

$$\text{minsup(Comb)} = \text{minsup0} + (\text{minsup1} - \text{minsup0})g(\text{Comb}).$$

CONCLUSION

The knowledge acquisition by means of generalization of association rules is one of the knowledge acquisition methods used in data mining. The benefit of this method consists in relative independence on high level programming solutions as for artificial intelligence tools. These tools require the active complicity of knowledge engineers and are often isolated from existing program support in related enterprises. The data mining upgrades the existing enterprise knowledge and is familiar to users from the agricultural area.

REFERENCES

AGRAWAL, R. – IMIELINSKI, T. – SWAMI, A.: Mining associations between sets if items in massive databases, Prof. Of the ACM-SIGMOD 1993. In: Int. Conf. on Management of Data, Washington D.C., May 1993, pp. 207–213.

AGRAWAL, R. – MANNILA, H. – SRIKANT, R. – TOIVONEN, H. – VERKAMO, A. I.: Fast discovery of association

rules. In: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press 1996.
BERKA, P.: Dobývání znalostí z databází. Academia 2003, pp. 103–113.
FAYYAD, U. – IRANI, K.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proc. Joint Conf. Artificial Intelligence IKCAI, 1993, pp. 1022–1027.
CHUNG, F. – LUI, CH.: A post-analysis framework for mining generalized association rules with multiple minimum supports. In: Proc. KDD-2000 Workshop on Post-Processing in Machine Learning and Data Mining, 2000.

Received for publication on January 14, 2008
Accepted for publication on March 3, 2008

VANÍČEK, J. – VOSTROVSKÝ, V. (Česká zemědělská univerzita, Fakulta provozně ekonomická, katedra informačního inženýrství, Praha, Česká republika):

Získávání znalostí ze zemědělských databází.

Scientia Agric. Bohem., *39*, 2008: 82–85.

Znalosti a informace hrají klíčovou roli v úspěšném provozování podnikatelských aktivit. Pro management znalostí lze mimo jiné využít i metodu získávání znalostí z databázových evidencí. Článek popisuje možnosti takovéhoto získávání znalostí ze zemědělských databází metodou zobecnění asociačních pravidel a tyto záležitosti demonstruje na problematice zemědělských ochranných prostředků. Takto získané znalosti mohou být cennou podporou například pro efektivní rozhodování pěstitelů při nákupu ochranných prostředků. Cílem předkládané práce je demonstrovat možnosti získávání dalších znalostí ze zemědělských databázových evidencí metodou zobecňování asociačních pravidel, přičemž tyto záležitosti jsou dokumentovány na příkladě metodické příručky pro ochranu rostlin vydávané Ministerstvem zemědělství ČR. Získávání znalostí z databázových evidencí lze přitom charakterizovat jako netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat (F a y y a d , I r a n i , 1993). Zdrojem pro tyto záležitosti mohou být existující relačně databázové evidence, ve kterých evidovaná data mají často podobu jednoduchých pravidel vyjadřujících příslušné znalosti.

Metoda získávání znalostí zobecňováním asociačních pravidel je jednou z mnoha dalších v repertoáru dobývání znalostí. Výhodou tohoto přístupu je poměrná nezávislost na programových řešeních vyššího stupně, jako jsou prostředky umělé inteligence, které vyžadují aktivní spoluúčast znalostního inženýra a velmi často jsou odtrženy od stávajícího programového vybavení daných podniků. Rovněž nepříznivě zde může i působit případná nechuť uživatelů k takovýmto novým, pro ně neznámým řešením. Opačně získávání znalostí z relačně databázových evidencí navazuje na to, co již v podniku funguje a pro uživatele není neznámým.

znalost; expertní systém; relační databázový systém; asociační pravidlo; získávání znalostí

Contact Address:

Prof. RNDr. Jiří V a n í č e k , CSc., Česká zemědělská univerzita v Praze, Fakulta provozně ekonomická, katedra informačního inženýrství, Kamýcká 1076, 165 21 Praha 6-Suchdol, Česká republika, tel.: +420 224 382 362, e-mail: vanicek@pef.czu.cz