Automatic on-line risk analysis of suspicious transactions

Automatická on-line analýza rizik podezřelých transakcí

P. Sobotka[1], I. Vrana[2]

[1] Komix s.r.o., Prague, Czech Republic
[2] Czech University of Life Sciences, Prague, Czech Republic

Abstract: Portion of the electronically processed agendas and transactions of all kinds constantly increases and also the volume and importance of the processed agendas grows. Therefore, the need for supervision and inspection of these transactions is also increasing. One possibility is to accomplish control e.g. with the use of specialised mining instruments. The main disadvantage of this approach is the fact that the discovered suspicious transaction was completed long ago and it is impossible to change it. In such a case, we need to inspect and evaluate transactions in the real time still before their completion. But the complexity of real-time analysis of transaction rapidly grows with the increasing set of aspects of this assessment. This paper describes the conception, architecture and possibilities to utilize a system which contains mainly the on-line risk analysis and simultaneously it enables an automatic adaptation by the means of utilizing conclusions from a feedback analysis of mining instruments of the data warehouse. This system is suitable e.g. for the automatic on-line analysis of risks of credit card payments, authenticity of the submitted projects (at university), the submitted tax or customs declarations, etc. Without any detriment to generality, we shall use the last mentioned application domain for explaining the system properties.

Key words: risk analysis, expert systems, composit system

Abstrakt: Vzhledem ke vzrůstajícímu podílu elektronicky zpracovávaných agend a transakcí všeho druhu a zároveň vzrůstajícímu objemu a významu operací uskutečňovaných touto cestou současně vzrůstá i potřeba provádět kontrolu těchto transakcí. Jednou z možností je provádět kontrolu zpětně např. pomocí specializovaných dolovacích nástrojů. Hlavní nevýhoda tohoto přístupu však spočívá v tom, že odhalená podezřelá transakce je většinou již dávno uzavřená a není možné ji měnit. V takovém případě je nutné ohodnocovat transakce v reálném čase ještě před jejich uzavřením. Složitost analýzy transakcí v reálném čase však prudce roste s rostoucí množinou hledisek, podle nichž se hodnocení provádí. Tento příspěvek pojednává o koncepci, architektuře a možnostech využití systému, který obsahuje především analýzu rizik v reálném čase, zároveň však umožňuje automatickou adaptaci prostřednictvím závěrů získaných zpětnou analýzou dolovacími nástroji z datového skladu. Systém je vhodný např. pro automatizovanou on-line analýzu rizikovosti plateb kreditní kartou, autentičnosti semestrálních projektů, podávaných celních deklarací apod. Bez újmy na obecnosti posledně jmenovanou aplikační doménu použijeme pro demonstraci vlastností systému.

Klíčová slova: analýza rizik, expertní systém, kompozitní systém

Recently the number of the electronically processed agendas of all kinds rapidly grows. With the increasing volume of electronically processed agendas, also the demands on the speed of transaction increased. This brings risk of their misusing, which requires performing of more and more detailed checking of all transactions.

There exist several methods supported by commercial products, which enable to accomplish an analysis of the transaction historical data (off-line analysis

Supported by the Ministry of Education, Youth and Sports of the Czech Republic (Grant No. MSM 604070904 – Information and knowledge support of strategic management and 2C06004 – Intelligent instruments for assessment of relevance of content of data and knowledge resources).

– e.g. Attoh-Okine, Ayyub 2004) and to identify the cases which deviate from other cases. The identified cases are then subject to a detailed checking. A large data-base as well as an extensive experience of the analysts applying these methods are the preconditions of their successful utilisation.

Based on this detailed checking, the majority of all identified anomal cases is usually considered risk-free, although they are somewhat non-standard from a certain viewpoint. The remaining transactions are then assessed as risky ones and some corrective actions are started if their character permits that.

The main benefit of the off-line risk analyses of the performed transactions is that they need no or little explicit a priori knowledge in order to detect the suspicious cases. This is possible by using various statistical data-mining methods with usually a large explored database. Off-line analysis has mostly an interactive iterative character.

A primary result of the off-line analysis is mainly an eventual set of transactions, which are suspicious from some aspects. Moreover, an advantage of the off-line analysis is its frequent capability to extract more general characteristics from the selected set of suspicious transactions, which can be later used as marcants for transactions classification. This secondary result of the off-line analyses, i.e. extraction of an explicit characteristic of suspicious transactions, can be a still more important result than the primary one (i.e. selecting a set of suspicious transactions) from the long term perspective. Methods without a priori information have a greater capability to discover new, unknown groups of risky transactions and their characteristics.

The greatest disadvantage of the off-line analyses is the fact that the suspicious transactions are closed long ago and it is often impossible to change them retroactively in the time of identifying these transactions as suspicious. Also in this case, the secondary result of the off-line analyses (i.e. extraction of explicit characteristics of suspicious transactions) gains a relatively greater importance.

Conducting an analysis at the moment when the transaction is taking place (so called on-line analysis – Latino, Latino 2006; Wasson 2005) is an alternative to the retroactive data analysis of historical transactions (off-line analysis).

Characteristics of the on-line analysis are to the great extent an opposite or a complement to the characteristics of the off-line analysis. Apparently, the most distinctive contrast is the capability of the on-line analysis to discover a suspicious transaction immediately in its course and to prevent its completing. This is the greatest advantage of the on-line analysis.

Somewhat less obvious is a palette of methods, which can be used in the on-line analysis and the properties of these methods. In the perspective of growing demands at the speed of conducting transactions, the on-line analysis is usually time limited to the order of seconds. Therefore, the on-line analysis cannot directly use methods without a priori knowledge based on mining of the historical data, because a generic property of all these methods is their high computational and time demand. Thus the on-line analysis should rely on the methods based on a large volume of very concrete explicit a priori knowledge. The consequent disadvantages are straightforward:

– High dependence of the analysis quality at the quality of a priori information.
– Incapability to discover new models of risk transactions.

As mentioned earlier, the on-line analysis is based on a priori information. Its quantity and quality significantly influences the resulting quality of the on-line analysis. But the quantity and quality of a priori information does not influence only the quality of the on-line analysis. It also has a strong influence on the computational complexity and therefore also time demands of the conducted analysis. The more extensive and detailed is the set of a priori information concerning risk aspects of individual cases, the more demanding this evaluation is.

Computational complexity of an on-line analysis does not depend only on the size of a priori information concerning assessment, but also on the character and size of the description of the evaluated transactions. The evaluation of some seemingly simple a priori information can have even an exponential asymptotic complexity. Thus it is clear that the implementation of the system for on-line analysis needs highly sophisticated evaluation methods, except of some very simple cases with a small set of simple a priori information having a small size of description of the evaluated transactions. But the "naïve" evaluation methods quickly degrade with the increasing load. Thus, they are not suitable for a real operation.

SYSTEM OF AUTOMATIC ON-LINE RISK ANALYSIS – SOLA

Motivation and requirements

Without any loss of generality, let us consider an on-line analysis of risks of the received custom declarations as a case study. This case study can serve as an example of the utilisation of this principle in the public and the state administration. Let us consider a need to replace the current outdated local system for the analysis of risks of custom declarations by a newer, modern and nation-wide system, which should be:

- Sufficiently permeable for managing daily operations on-line.
- Open to a future extension for the additional processed agendas.
- Highly flexible in the sense of creation and administration of a priori information for the evaluation of risks.

The considered system should process in average about twenty thousands of declarations daily. About 80% out of them should be processed during 6 hours. The peak load is thus about 45 declarations per minute. With respect to the declarations complexity, we therefore need to design such system architecture as to be able to evaluate the received declarations in seconds.

The mentioned volumes of declarations should be evaluated against a large set of a priori information. The size of the a priori information set is assumed of the order of thousands "business rules" of various complexity. Higher level rules are called risk profiles. During the execution of each incoming declaration (or generally of any document), this document should be compared with all rules from the set of a priori information. It should be checked at each rule from this set whether the processed document complies with the rule or not.

At the same time, an architectonic design should consider that the volume of the processed documents will significantly grow in future as a consequence of the planned adding of further types of processed agendas. The system should therefore have such architecture as to be capable of the broadest scaling, i.e. increasing its real performance by adding independent computing units (processors or computers).

A very important requirement put on the newly built system is to be in the maximum open with respect to administration, maintenance and extending a set of a priori information. It should be possible to change or modify the set of a priori information to the largest extent without any necessity of the additional application programming. This requirement follows from the fact that models of risky behaviour could rapidly change and the system should quickly adapt to these new situations. A standard IT system development cycle is very slow from this perspective. Another important reason for the user-administrated set of a priori information is the fact that the carriers of knowledge about the models of suspicious behaviour are the users, not the developers. Last but not least, the models of risky behaviour are the most valuable and the most secret part of the system and therefore the least number of people should be acquainted with their content.

Conception

The above mentioned requirements could be satisfied by the SOLA system (Vrana 2006). It is an expert system the conception of which comes from its primary purpose to perform an on-line analysis (the risk analysis of custom declarations in our case). As mentioned in the introduction, the on-line analysis is supported by the set of a priori information against which the incoming declarations are evaluated.

A priori information is in the system referred as risk profiles. They contain a full model of risky situations and they generally consist of:

- detection part
- reactive part.

The detection part of a risk profile describes an assumed model of declaration, which is considered risky. It should be checked, whether the processed declaration corresponds to some of the models defined by the detection parts of the given risk profiles. The processed declarations are therefore compared with detection parts of all individual risk profiles. The description of the risky declaration model has a form of a subset of partial characteristics, which can be combined to an arbitrarily complex logical term. If the given logical term has the value TRUE at the processed declaration, it means that the declaration corresponds to the model defined by the corresponding risk profile.

The reactive part of the risk profile defines what should happen (how the system should react) in the case, when the processed declaration corresponds to the model defined by the corresponding profile. The reactive part of the risk profile could have the form of one or several simple actions (e.g. generating an informative message, sending an SMS, e-mail, ordering an inspection, etc.) in a simple case. However, the reactive part could be described also by a very complex algorithm in some cases.

The SOLA system performed an on-line analysis of risks of the processed custom declaration in this case study. The characteristics, i.e. the advantages and disadvantages were briefly described in the introduction. But the SOLA system is not constrained to the pure on-line analysis as described in the introduction, i.e. it is not based only on the static a priori information = risk profiles. The SOLA introduces the connection of a pure on-line analysis with an off-line analysis in order to reduce the main disadvantages of the pure on-line analysis. This connection extends the capability of the on-line analysis by the ability of an automatic adaptation of behaviour with respect to the continuously changing real world characteristics, e.g. fraud behaviour.

The ability of an automatic adaptation of the SOLA system strongly extends its possibility compared to the pure on-line and off-line methods of analysis. An automatic adaptation is based on the operation of the abstract part of the model of the expected risk situation, which is more or less static and manually administered, from the specific part of the model, which has substantially greater time-dynamics and is administered automatically. This automatic adaptation can be used both in the detection as well as in the reactive part of the risk profile.

Let us consider a trivial case of the risk profile, which would evaluate declarations according to the specific costs of goods. Let us further consider that the declaration should be treated as suspicious and the further investigation is needed, if the declared cost considerably differs from the "prevailing" price of the given commodity. In this trivial case, the abstract portion of the detection part can have the form of the comparison of the declared specific price with the average price of the given commodity. The required reaction depending on the detected state would then be described in the reactive part (e.g. strongly over average, average, strongly below average). In this case, the detection part would have the concrete form of the list of the average specific costs of the individual commodities. This list should be regularly automatically updated from historical data.

It is clear that such a risk profile has an unlimited time validity. The principle itself (i.e. comparison of the declared and "usual" price) does not change, but the list of the average "usual" costs is changing. This change is automatically calculated from historical data and it enables to react flexibly e.g. to seasonal deviations of some commodities and to preserve the required accuracy of the on-line analysis during the whole lifecycle of the given profile.

Obviously, one can speculate what does the "usual" price mean, how is this price calculated and how to determine to which extent the declared price corresponds to the "usual" price. The SOLA system is so flexible that it puts all these aspects fully into the hands of the user – the designer of the risk profiles. The user has fully under control the way how the parameters are calculated ("usual" price) and also the way of comparison of the declared and calculated parameter. In a naive case then the "usual" price could be represented by the arithmetic mean and the comparison of declared value could be accomplished by the simple operators "<", "=" and ">". Such a trivial model, however, would not be of the required accuracy. But regardless of the complexity of the model, the designer can always materialize it by the user means. A simple arithmetic mean can be substituted e.g. by a more complex statistical function or an arbitrary heuristic calculation. Also the way of comparison of the declared and calculated parameter is fully in the hands of the user. A rich apparatus of the predefined general operators is available. If yet some special function or operator would be needed, The SOLA can extend the assortment of offered operators limitlessly, according to the users requirements. Generating of the parameters for risk profiles need not necessarily be a part of the SOLA system and therefore arbitrary available means could be used including the specialised statistical and mining programmes.

Architecture

The architecture of the SOLA system is schematically illustrated in Figure 1. The right part of the picture is the most important from the operational point of view. The transaction begins in the moment when the transaction indicator provides the triggering information. In our considered example of checking custom declarations, a subject submits a custom declaration, the customs official receives it and the client application transmits it for the risk assessment (e.g. in the XML shape). The SOLA system has a central character, i.e. all received declarations are sent for assessment to the centre and the response travels back.

The evaluation of the risk amount takes place according to the preset risk profiles. Two types of risk profiles occur in the SOLA system:
– tactical profiles
– strategic profiles.

The mission of the tactical profiles (also called blocational profiles) is to describe as simply as possible a relative simple, concrete risk situation model. In the considered example, it typically could be the request to stop or inspect the declared shipment, etc. Tactical risk profiles could be defined and administered exclusively manually by a rather broad users community e.g. by the means of the WWW application with a simple form-interface, where the

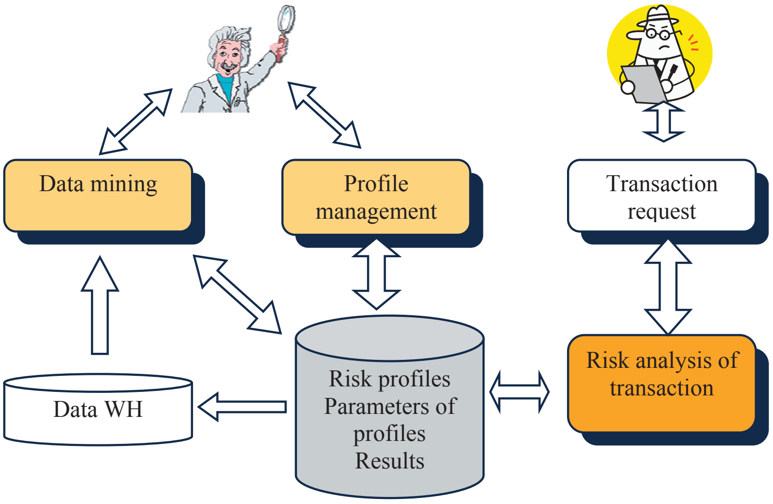


Figure 1. Architecture of the SOLA system

individual users fill-in the items, which in the incoming declarations should be searched for. In this case, a link with the off-line analysis is only indirect in the way that the results of analysis could be recorded in the shape of risk profiles and based on that, the on-line analysis was performed.

Strategic profiles have a different mission, character and philosophy. In difference from tactical profiles, the purpose of which is simplicity, strategic profiles give a full freedom to their designer and also provide a very broad palette of instruments to formalise even very complex models of risk situations, both in the detection and reactive part.

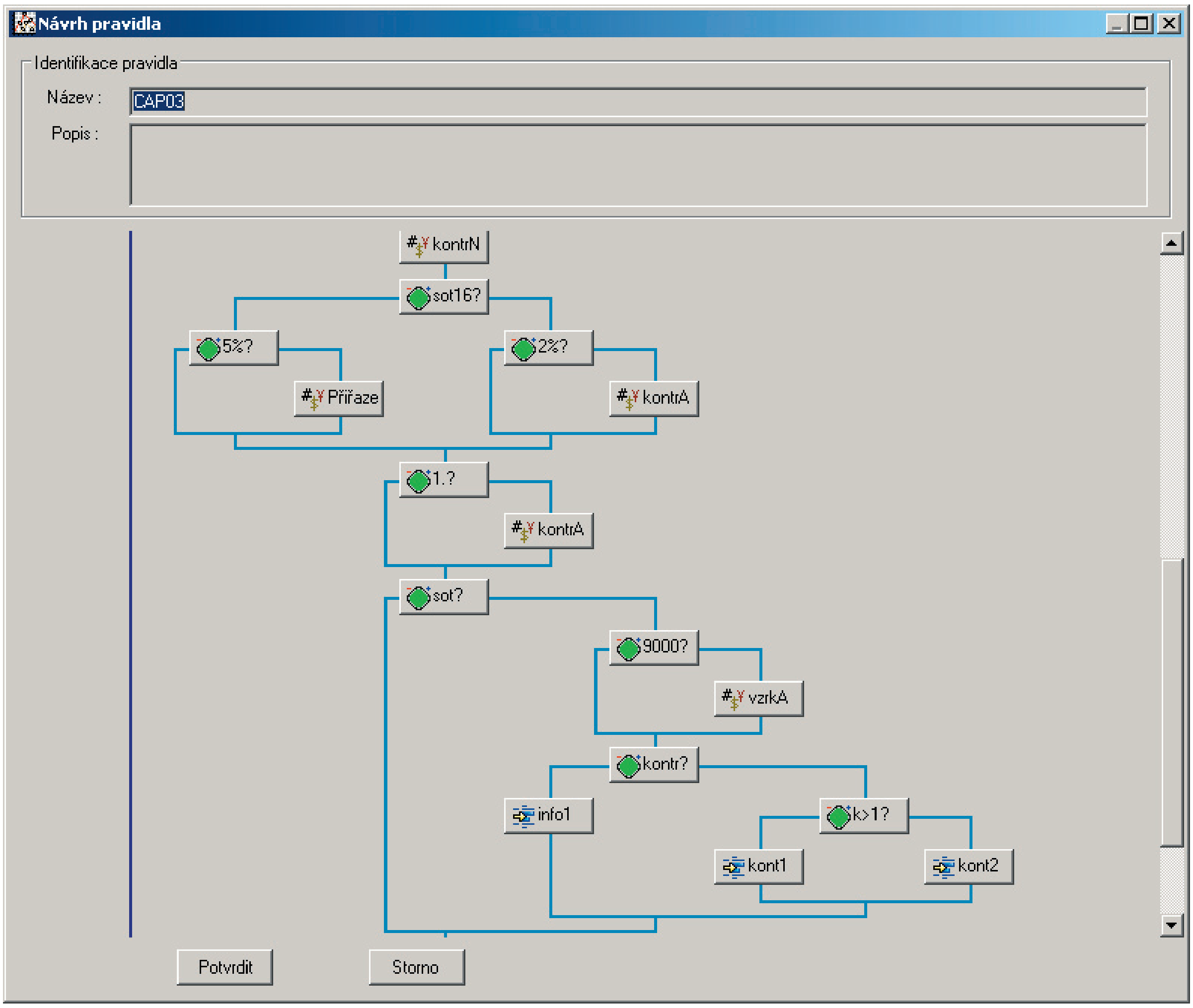
We can create a strategic profile with the functionality of a tactical one in the over-simplified case, but this is not the purpose of the strategic profiles. All predefined operators and functions could be utilised in the strategic profiles to create the required functionality. Moreover, these operators and functions could be combined to more complex formulas according to the given syntax rules or still further operators and functions can be added.

In contrast to tactical profiles, creation of strategic profiles needs very knowledgeable designers, who must perfectly master the businesslike nature of the modelled problem and also they should have some skills in algorithmic. Strategic profiles offer a general development environment with the possibility to use items of an imperative and a declarative programming. Particularly, a declarative programming represents a very powerful instrument for formalisation of the required model without a need to imperatively define the way of evaluation. Conversely, if there was a need, the strategic profile could contain also a set of complex algorithms. Figure 2 graphically depicts building of an algorithm of a strategic profile in the system for the custom declarations risk analysis.

Risk profiles primarily work with information coming from the analyzed transaction, from the assessed declaration in the considered example. Strategic profiles could moreover utilize also the results of other strategic profiles and create further, still more complex models. Besides the results of other strategic profiles, strategic profiles can work also with the so called external parameters and utilise a link between the off-line analysis and between abstractions from concrete values (see Conception).

A high degree of generality and universality offered by the SOLA system by the means of strategic profiles has, of course, also its negative consequences. These are high demands at the professional skills of the designers of these profiles. A high degree of freedom does not prevent designers to build such profiles, which have a high computational complexity or are logically incorrect, although they are correct syntactically. To support building and debugging of profiles, the SOLA system provides the possibility to test the created profiles against the historical data.

Risk profiles, their external parameters and statistics reside in a central system database. The operators could administer profiles with the use of two applications: one for tactical, the second for strategic profiles. The operators draw from the off-line analyses of the historical data stored in the data warehouse. They can utilise various analytical tools which, equally as the data warehouse, are not the components of the SOLA system. There exists still one direct link between the data warehouse (off-line analysis of the stored data) and risk profiles of the SOLA system.



Návrh pravidla

Identifikace pravidla

Název : CAP03

Popis :

Figure 2. Algorithm within a strategic profile

It has the form of automatic calculations of external parameters for some strategic risk profiles.

Performance

As mentioned in the Introduction, the required throughput of the system (of the system of risk analysis of custom declarations in our example) is of the order of tens up to hundreds of evaluated declarations per minute with thousands of profiles of various complexity. It is clear that no "naive" assessment methods could be used at the expected rate of the incoming declarations where all profiles would be repeatedly read and interpreted for every incoming declaration. If the profiles were stored in a database, then they could not be even read (not to speak about the interpretation of their complex constructions and difficult calculations) in the given time constraint.

An approach of the off-line translation of the created risk profiles into the shape of a set of elementary rules was selected for the implementation of the core of the SOLA system. The obtained rules are then interpreted by an inference machine based on a well-proven RETE algorithm invented in the NASA by Forgy (1982). The off-line translation itself prepares the profiles for interpretation. The repeated gradual reading of the profile content in every assessment is eliminated. An inference algorithm provides a very high efficiency of evaluation with respect to the size of the set of rules.

The system modul which cares about the evaluation of the processed transactions is separated from the module for risk administration and can be operated on a separated hardware. There are several reasons supporting this conception, but the following three belong to the major ones:

– load distribution,

– scalability – the assessment module can be installed on further hardware in order to increase the performance,

– stability and security – the assessment is not influenced by the profiles administration in any way.

Used technologies

The SOLA system is inherently independent on the used technology, only its assessment core – the inference machine - should be implemented with the use of the most powerful technology. In the case of our example of the custom declaration analysis, the following technologies are used:

- Data exchange (incoming declarations, results of evaluation): XML
- Profiles administration: .NET
- Client applications: .NET
- Critical parts: C
- Database: MS SQL

PRESENT AND FUTURE

The SOLA system was tested for the risk analysis of custom declarations during its many-months routine operation. Its expected performance, robustness and also flexibility were fully verified and justified during this trial. The system even by far exceeded the required parameters in some respects.

The SOLA system was developed from the scratch and it provides radically new possibilities in conducting the risk analysis of processed declarations. A perfect mastering of all possibilities offered by the SOLA and also discovering of further needs will require a longer deployment of the system in further application domains. The conception of the SOLA provides the possibility to be utilized in the analysis of any other on-line agenda, which requires a high performance and flexibility from the point of view of behaviour definition, e.g. agendas related to the the fraud detection in banks and offices, with problems in protocols detection, in logistics, the document flow, etc.

From the technological point of view, in future it could be possible to consider an upgrade of the tools for description and evaluation of risk profiles, the tools for ambiguity handling (e.g. utilizing fuzzy logic), which could add still more robustness to strategic profiles and enhance the natural understanding. The process of tuning of the strategic profiles can also be enhanced, for which only a basic support is available now. Also the support of distributionality can be further enhanced by interlinking of several instances of the system in order to share the risk profile definitions, their parameters and characteristics.

CONCLUSION

By its conception and the offered tools, the SOLA system represents a very innovative step towards an automated on-line analysis of the suspicious transactions risk. The exceptionality of the system lies in the degree to which the users can modify the system behaviour by the means of the strategic profiles and the possibility to match this system to the new knowledge and needs without programming interventions into the system. Further, it is exceptional because of the class of tools available through strategic profiles, because of the automatic enhancement of the on-line analysis by the results of then off-line analysis, and, last but not least, because of the effectiveness of processing and openness towards changes in processing of other new agendas.

REFERENCES

Attoh-Okine N.O., Ayyub B.M. (eds) (2004): Applied Research in Uncertainty Modelling and Analysis. International Series in Intelligent Technologies. Springer Science, Germany.

Forgy C. (1982): Rete: A fast algorithm for the many pattern/many object pattern match problem. Artificial Intelligence, 19 (1): 17–37.

Latino R.J., Latino K.C. (2006): Root Cause Analysis: Improving Performance for Bottom-Line Results. 3rd Edition, Taylor & Francis, London.

Vrana J. (2006): Kombinovaná automatizovaná analýza rizik celních deklarací. DATAKON 2006, pp. 295–304.

Wasson C.S. (2005): System Analysis, Design and Development: Concepts, Principles and Practices. Wiley Series in Systems Engineering and Management.

Arrived on 25th September 2007

Contact address:

Ivan Vrana, Czech University of Life Sciences Prague, Kamýcká 129, 16951 Prague 6-Suchdol, Czech Republic
e-mail: vrana@pef.czu.cz