SELECTION WITH THE HELP OF DATA MINING

LÁSZLÓ KOVÁCS
Department of Information Technology, University of Miskolc
H-3515 Miskolc, Hungary
kovacs@iit.uni-miskolc.hu

MARIANNA LIZÁK
Department of Human Resource, University of Miskolc
H-3515 Miskolc, Hungary
alkmlm@iit.uni-miskolc.hu

GÁBOR KOLCZA
Department of Information Technology, University of Miskolc
H-3515 MISKOLC, Hungary
kolcza@iit.uni-miskolc.hu

[Received October 2004 and accepted February 2005]

Abstract. One of the most important resources of a company is the human resource. An economic organisation has to take care of procurement and human resource management. At procurement it is decided in the process of selection which candidate will be given a position-offer. A great importance is set on this decision during the life of the company. Data mining ensures an adequate background for making well-grounded decision. This article is dealing with the methods and technics that make this process faster, more efficient and more reliable.

1. INTRODUCTION

Nowadays it is a widely accepted fact that the most important resource of a company is the human resource. One precondition of an efficiently working and profitable company is that it has to have an adequate human resources strategy. This includes the procurement (recruiting, selecting and launching), management (manpower development, performance appraisal and career planning) and the 'drain' (reducing staff, retirements, etc.) of the human resources. From the cost effectiveness point of view it is equally important that the man-power should be well-skilled and performance-orientated, thereby making profitable the organisation. It is an important point of view as well that the possible least cost is to be spent for manpower-development and recruiting, because costs can be saved with this (Figure 1). One of the important decisions for human resources managers

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

is that these two mutually exclusive factors should be optimized. Thus bring out the maximum profit and performance from the organization.

One key-task of human resources strategy is the selection. The selection is a filter which has the task to pick out from the applicants the candidates who best fulfil the requirements of the particular positions and trustworthily classify them. The selection process has a considerable responsibility, as in this phase of recruiting of staff we arrive at a decision about who will be the employee of the organization thereby having an influence on the performance and profit of the company. Selection as other activities related to human resources goes with time- and energy-cost which has no direct effect on profit. In companies the need has come to the front to reduce somehow the expenses of such tasks. With the improvement of information technology and mathematical methods and the appearance of data mining the opportunity is given to reduce these costs without reducing the energy devoted to human resources. [2]

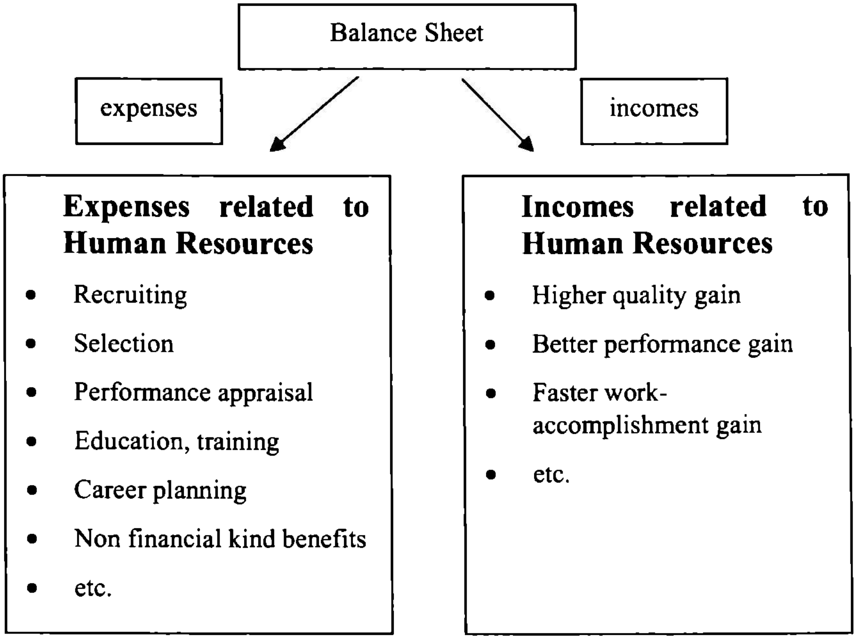

Figure 1: The optimization of the expenses and incomes related to human resources

2. OVERVIEW OF DATA MINING CONCEPTS

Because of the increasing competition among companies and the extreme consumer expectations the companies have to carry out continuous developments and analysis in order that they can keep their market position. In order to do this

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

they have to process huge amount of data. To solve this problem the data mining has arisen, which became popular in the business world in the 90's. Data mining is more and more widespread in information technology and business. Data mining gives opportunity for the companies to handle the enormous amount of data with adequate efficiency and from which they can distil useful information in order to achieve greater profit. According to the definition of the SAS Institute the task of data mining can be composed as follows: 'the process of selecting, discovering and modelling huge amount of data, which purpose is to find correlations and patterns of data, that could have not seen in advance, in order to gain business advantage.'

In this approach it is important to make difference between the data that can be found in various files or databases in unprocessed form, that are actually useless, and the information that is distilled from data and used to gain business advantage.

2.1. The application of data mining

Nowadays there are lots of application of data mining, such as: health care, direct marketing, commerce, risk-analysis, logistics, telecommunication, transportation, decision-support and human resource management. In commerce with the obtained information it is used to increase customer satisfaction, turnover and profit. In logistics with the integration of data mining the efficiency and promptness of distribution can be improved.

It is more and more widespread in large enterprises that data mining is used in human resources management, since by this means considerable cost-saving may be achieved. Also at the Dutch airways KLM as well data mining was applied in favour of cost reduction. One determinant costfactor of airways the cost of pilots, i.e. the human resources. This system was necessary because serious career-planning is carried out in the company. For the pilots the company assures continuous progress, i.e. there is a continuous labour fluctuation inside the company. The problem is that they do not know what kind of jobs the pilots will apply for. To help solving this problem they apply the software called CAPTAINS, which is developed for the Dutch airways company KLM, and which uses data mining as well. This is a planning and optimizing software in which machine-learning methods are used as well.

This system makes it possible to foretell what kind of positions the pilots are going to apply for and thereby the right number of pilots will be employed, neither too many, nor too few. KLM with applying this human resources planning software could reduce the human resource management cost by 2 per cent and the cost of investing this system returned in one year. [4]

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

Data mining can be used in different fields. The different fields however need different methods, which are discussed in details in the following chapter. [5][6]

2.2. Data mining methods

The term data mining includes all methods and techniques that can be used for discovering new and relevant rules and dependencies from a huge amount of row data. The algorithms are based usally on statistical and heuristic methods. The data mining applications are complex systems as they require the adoption of the general methods to the special problem. The parametrization of the methods and the interpretation of the results requires expert users. The data mining systems are useful tools only in the hand of good experts.

From the logical and functional viewpoint, the process of data mining can be divided into the following phases:

- The first step is the analysis of the required information and the available information sources.
- The analysis of data format of the data sources. The quality of the data sources should also be examined.
- Data from different sources should be transfered to a common data stage. The data transport includes the extraction, the transportation and the format conversion steps.
- The next step is the data integration on the staging area. This includes the development of the common data schema and the discovery of the inconsistencies among the data from different sources.
- To increase the efficiency of computation, the amount of data to be proccessed should be decreased with some kind of data reduction method.
- A pre-analysis phase with some kind of OLAP or statistical tool. These tools help to illustrate the overall behavior of the data set without any deep analysis. This helps to localize the problem areas.
- The definition of the concrete goals of the analysis. The selection of the data mining tasks best fitting to the investigated problem area is an important step done by experts.
- Based on the task, the selection of the data mining algorithm and method best fitting to the investigated task is also an important step done by experts.

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

- The definition of the parameters and the constraints related to the selected data mining method.
- Running the tests.
- Interpretation of the results, definition of the discovered rules and dependencies for non-expert users.
- The evaluation of the method, how correct is the generated result.

The algorithms of the data mining tools are based on some well-known mathematical methods. In the practice, the following methods are usually applied for rule discovery problems:

- Discovery of association rules. The goal is to determine which objects implicate wich other objects. The method determines first which objects occur often together and the direction of the implication rule is calculated next.
- Classification. The training objects are assigned to some predefined classes. The method discovers the hidden dependencies among the object parameters and the class labels assigned to the objects.
- Clustering. Only the training objects are known without any predefined class labels. The method determines the groups of similar objects based on their attributes. The number of groups is usually not known a priory.
- Detection of typical event chains. The method discovers the ordering of the events and the most possible chain segments.
- Deviation analysis. First the group of objects with average behaviour are determined then the outlier objects are discovered. The calculation is based on the object attributes.
- Nearest neighbour search: The training set includes a large amount of samples. The method is aimed at determining the most similar objects to a query object. [6]

In our investigation, the classification method is the appropriate method. In classification processes, it is assumed that the patterns are stochastically independent. A d-dimensional pattern vector is denoted by $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \in \mathbf{R}^n$ Every pattern vector is associated with a class c_j, where the total number of class is m. Thus, a classifier can be regarded as a function

$$g(\mathbf{x}) : \mathbf{R}^n \rightarrow \{c_1,\ldots,c_m\}. \tag{1}$$

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

The optimal classification function is aimed at minimizing the misclassification risk. The R risk can be measured by an appropriate cost value. The risk value depends on the probability of the different classes and on the misclassification cost of the classes.

$$R(g(\mathbf{x}) \mid \mathbf{x}) = \Sigma_{\mathrm{cj}}\, b(g(\mathbf{x}) \rightarrow c_j) P(c_j \mid x), \tag{2}$$

where $P(c_j \mid \mathbf{x})$ denotes the conditional probability of c_j for the pattern vector $\mathbf{x}$ and $b(c_i \rightarrow c_j)$ denotes the cost value of deciding in favour of c_i instead of the correct class c_j. The b cost function has usually the following simplified form:

$$\begin{aligned} b(c_i \rightarrow c_j) = {} & 0, \;\text{ if } c_i = c_j \text{ and} \\ & 1, \;\text{ if } c_i \neq c_j. \end{aligned} \tag{3}$$

Using this kind of b function, the misclassification error value can be given by

$$R(g(\mathbf{x}) \mid \mathbf{x}) = \Sigma_{\,g(\mathbf{x}) \neq \mathrm{cj}} P(c_j \mid \mathbf{x}). \tag{4}$$

The optimal classification function minimizes the $R(g(\mathbf{x}) \mid \mathbf{x})$ value. As

$$\Sigma_{\mathrm{cj}}\, P(c_j \mid \mathbf{x}) = 1 \tag{5}$$

thus if

$$P(g(\mathbf{x}) \mid \mathbf{x}) \rightarrow max \tag{6}$$

then the

$$R(g(\mathbf{x}) \mid \mathbf{x}) \tag{7}$$

has a minimal value. The decision rule which minimizes the average risk is the Bayes rule which assigns the **x** pattern vector to the class that has the greatest probability for **x**.

The Bayes classifier that minimizes the misclassification error is defined by

$$qB(\mathbf{x}) = \operatorname{argmax} q_j(\mathbf{x}), \tag{8}$$

where q_j is the a posteriori probability of class *j* at pattern **x**:

$$q_j(x) = P(c_j \mid \mathbf{x}). \tag{9}$$

The misclassification cost is equal to

$$R(g(\mathbf{x}) \mid \mathbf{x}) = 1 - qB(\mathbf{x}). \tag{10}$$

The lower is the $qB(\mathbf{x})$ value the greater is this cost. The greatest cost is yielded if every class has the same probability for the pattern vector **x**:

$$q_j(x) = 1/m\,, \tag{11}$$

in this case

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

$$R(g(\mathbf{x}) \mid \mathbf{x}) = 1 \quad 1/m \tag{12}$$

The lowest misclassification value is equal to 0. This occurs if only one class has a nonzero probability for the pattern vector, i.e. $qB(\mathbf{x}) = 1$. [6][7]

3. THE PROCESS OF SELECTION

Data mining can be used in the process of selection independently of the fact that it is carried out with inside, outside, innovative or conservative strategy. Figure 2 shows the process of selection. It can be seen that selection is not an easy task. In function of the size of the company and the type of the vacancies this process can be simpler or more complex, thus some examination may be cancelled or even may be expanded.

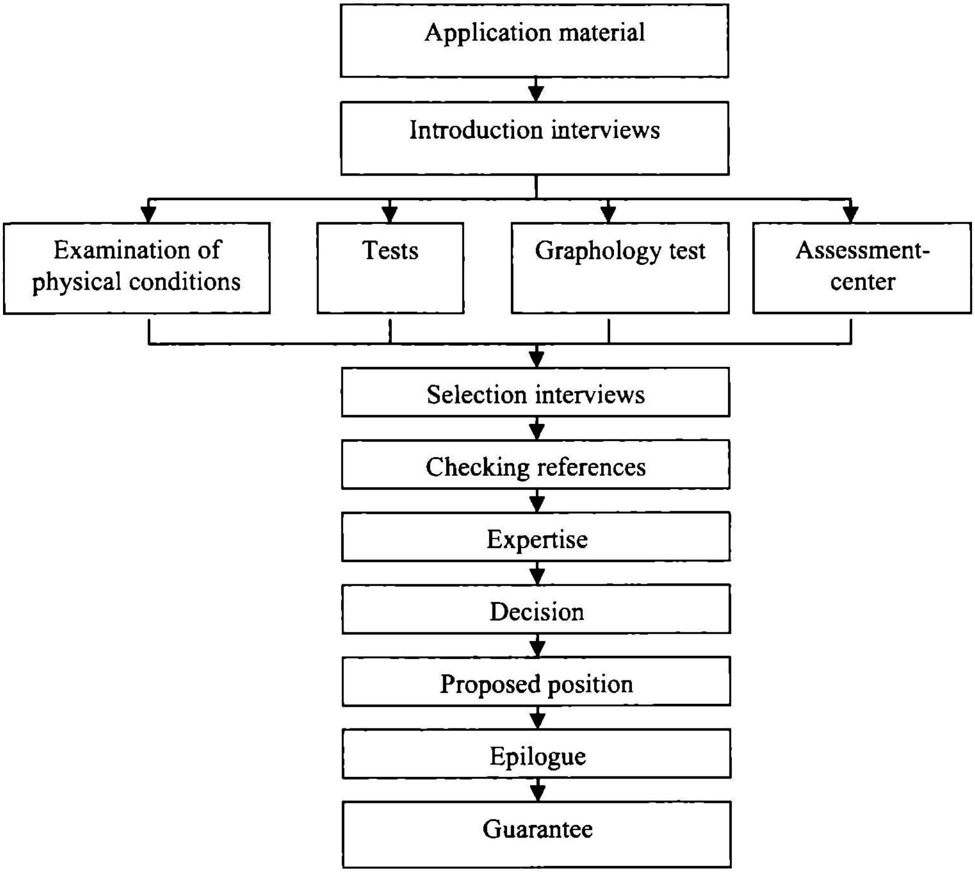

Figure 2: The process of selection

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

The complexity of the selection does not depend only on the size of the company but on the importance of the vacancy.

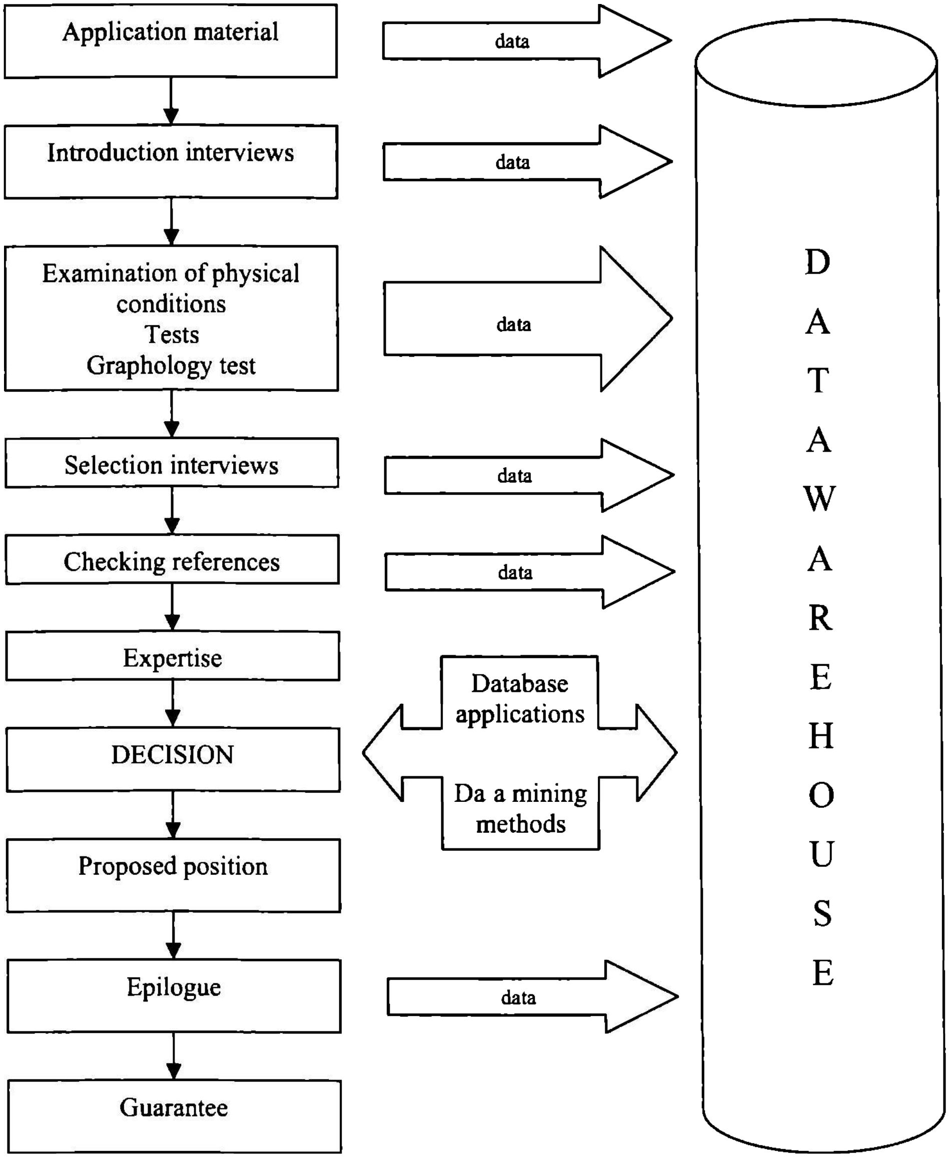

Figure 3: The process of selection with applied data mining

In order to be able to use data mining in selection we need a datawarehouse filled with data. In the age of Internet it does not make difficulties to store cv-s, motivation letters and other test results about a candidate in a database, as most company has an application form available on Internet.If a candidate fills it automatically gets in a database. On Figure 3 we can see how selection can be made effective with data mining. [1][2]

Nowadays enterprises record almost everything on computers. Thus it cannot be a problem to store useful information in datawarehouses.

In every step of the selection we record data in the datawarehouse. By Inmon „a Data Warehouse is a subject-oriented, integrated, non-volatile, and time variant collection of data in support of managements decisions".The datawarehouses give a possibility for storing a huge quantity of data in a suitable structure and for ensuring the quick performing of putting questions on the data.Instead of datawarehouse traditional databases or files can be used, though it has the disadvantage that the access of data is slower and the storage of same amount of data consumes more space. At application material we can store the qualifications, further purposes, marital status etc. of the candidate. At introduction interviews the personal impression of the examiner, the results of the candidate in various examinations, personality-, IQ- and other tests can be stored. In the selection interview we can get information about the candidate's motivation and characteristics.

The most important thing is to store data in the datawarehouse in the epilogue phase, because in this step turns out how well the new employee is approved. This ensures the feedback which helps to find out which attributes the 'good' and 'best' manpower possess.

It is important to have as much data as possible, because the more data we have the more efficient data mining can help in selection to classify candidates and make better decision. [8][9]

4. SELECTION WITH THE HELP OF DATA MINING AT THE DEPARTMENT-STORE OF TESCO, MISKOLC

The TESCO Global Inc. makes retailing by selling 70000 different types of products. To carry out this an adequate number of employees with the necessary qualification is needed. Currently about 300 employees work in the TESCO in Miskolc. According to this an effective staff-procurement strategy has been elaborated.

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

If there is some vacancy, the corresponding candidates are selected primarily from the database, maintained by TESCO. If in that database an adequate candidate is not found, which has a quite small chance, because currently there are 6000 pieces of application given in, the position is advertised in the press. For these advertisements application forms made by TESCO are accepted because in these forms all necessary data are asked for. The first step of filtering is based on these data. Each application material is provided with a serial number and every applicant has to sign a paper suitably to the relevant passage of the Data Protection Act that the data of the applicant is going to be deleted in half a year. The application materials are stored in Excel files, and it is classified into a position to which it is potentially suitable. After one month of the registration of the applicant's data a respond is sent to the applicant which informs about having been registered in TESCO database and as soon as there is a job the applicant will be contacted.

In case of a vacancy, only the application materials will be found that fulfil the requirements of qualifications and expertise. These filtered data is given to the head of the department who arranges a suitable time with the applicants and selects from the applicants by a personal interview. In case of every position the interview to which the managers are prepared in a training is enacted similarly.

The data needed to data mining are stored in Excel spreadsheets, which is not the most effective method as the structure of the files could not make possible the fast access, but it is one solution to store and access data.

In TESCO department-stores there are various performance classification levels into which every employee is classified once in a half year. These are registered in files as well. The method of association makes it possible to determine that mostly what kind of attributes from the application form are associated to the excellent and good performance. With the help of this method a rule-based deduction system can be made. In this case this can be regarded as a learning-algorithm, as it filters the most adequate by the experiments came from the existing data. With the association method relations can be discovered. These relations, however, usually are ambiguous, stochastic-featured. The closeness of stochastic relations can be analysed with coefficients and corellation- and regression-calculation used in statistics. The closeness of the relationships can be analyzed simply, clearly and effectively by means of the association coefficients therefore I will investigate this method in the future.

With the Yule association coefficient only relations between alternative criterions can be analysed. Primarily we would like to analyse the relation between the applicant's data and later performance. This could be e.g. the relation between the

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

received score for communication skill or other skills during the interview and the later performance evaluation. This coefficient could only be used in this case if we put aside the scalability of an attribute and we would only determine if the applicant has or has not the specified skill. On the one part this would not be the reality and on the other part this would worsen the efficiency of the software.

With the use of Csuprov association coefficient there is a possibility to analyse the attributes that are not alternative. In the first step the necessary data should be inserted in a contingency-table (Figure 4).

A\B	B_1	B_2		B_j		B_t	Σ
A_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1t}	$f_{1.}$
A_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2t}	$f_{2.}$
...							...
...	...	...	...	...	...	...	...
A_i	f_{i1}	$f_{.2}$	...	f_{ij}	...	f_{it}	$f_{i.}$
...							...
...	...	...	...	...	...	...	...
A_s	f_{s1}	f_{s2}		f_{si}		f_{st}	$f_{s.}$
Σ	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.t}$	n

Figure 4: Contingency-table

With the use of this association coefficient the relation between two attributes of the applicants can be analysed. Thus e.g. we can determine how strong the relation between qualification (elementary, secondary, higher) and communication skill or customer-orientation. In the TESCO supermarkets the ranking of performance levels of employees is indicated by different letters in a given field as e.g. in the fields of the communication relationship and customer-service. In a given field the excellent performance is indicated by A and B, the good one – by C, the weak performance is indicated by E and F; D is used for indicating the developing performance and X for indicating the performance that cannot be evaluated. According to it the relationship between these two attributes can be determined in the following Table. Figure 5 shows this.

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

Customer-orientation / qualification	A	B	C	D	E	F	X	Σ
Elementary	f_{11}	f_{12}	…		…			$f_{1.}$
Secondary	…	…	…		…			$f_{2.}$
Higher	f_{31}						f_{37}	…
Σ	f_{1}	$f_{.2}$	…	$f_{.j}$	…	$f_{.t}$		n

Figure 5: Analysis of relation between performance and qualification

Based on the contingency-table f_{ij}* (frequency assumed for independence) can be determined:

$$f_{ij}^{*} = \frac{f_{i.} * f_{.j}}{n} \tag{13}$$

From which χ^2 can be calculated by the following equation:

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(f_{ij} - f_{ij}^{*})^2}{f_{ij}^{*}} \tag{14}$$

The Csuprov association index can be calculated by the following formula:

$$T = \sqrt{\frac{\chi^2}{n * \sqrt{s-1} * \sqrt{t-1}}} \tag{15}$$

As the number of knowledge variants are not equal, hence instead of Csuprov coefficient the Cramer index is used, which can be calculated as follows:

$$C = \frac{T}{T_{\max}} \tag{16}$$

where

$$T_{\max} = \sqrt[4]{\frac{s-1}{t-1}}. \tag{17}$$

With the help of Csuprov association coefficient we can analyse how strong the relation between the attributes of an applicant. Besides this naturally all possible reasonable attribute-pairs should be examined.

With the combination of assotiation and similarity methods there is a possibility to classify the applicants filtered by the rules created with the association method by selecting the applicants whose application material resembles best in attributes to

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

the application material of excellent and good performance employees. It presents itself that clustering belongs to these two methods. With the help of this the classification of applicants can be carried out with higher confidence, as not only the data of individual employees are analysed but also the group of excellent and good performance employees and the rules are determined by these criteria as well.

But we should think about that generally the performance of the employees is non-uniform. Sometimes it comes to a rise another time to a fall. A generally excellent worker can present poor performance sometimes. If we take the previous combination example then it can be seen that it cannot give adequate confidence, as the created rules are strongly influenced by the performance of the individual eemployees at any given time. This problem can be eliminated by the method of prediction. This is a statistical method as well that takes account of the trend of performance, the seasonal variation and turns in performance as well. In a word if this method is integrated in our existing system then the variation in performance can be considered and our rules are not based only a given month's performance.

At analysing time series in the first step we should determine if it is additive or multiplicative. The time series analysed by us are in any case additive, as the employee is fired in case of too big and increasing variations in performance. In case of additive time series we apply the following formula:

$$y_{ij} = \hat{y}_t + S_j + c + v_t. \quad (18)$$

Where '$\hat{y}_t$' means the trend-value of the time series at a given 't' moment, 'S_j' means the seasonal variation, 'c' is the random effect and 'v_t' is the turns in performance. With this formula we can calculate the probable performance at a later date. This value is determined by four factors. The first factor is the '$\hat{y}_t$' trend , which determines the basic direction of the time series. Because of the nature of the data here we will calculate with linear trend. The linear trend-function can be determined by the following formula:

$$\hat{y}_t = b_0 + b_1 * t, \quad (19)$$

where

$$b_0 = \frac{\sum y}{n}, \qquad b_1 = \frac{\sum t * y}{\sum t^2} \quad (20)$$

The next factor determines in what direction and what extent the seasonal variation and the period deviates the data of the time series from the basic direction. Subtracting the trend value from the original value the seasonal variation can be determined for each quarter.

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

We do not study the cyclical swings, as for such a long time data are not available.

However we can calculate with the random factor – after having calculated the seasonal variations (S_j) – by the following formula:

$$y_{ij} - \hat{y}_{ij} - S_j = v_{ij}. \qquad (21)$$

Using the method of prediction alone there is a possibility to reduce or eliminate the performance-swing of the employees, as with the help of the method we can recognize the events and circumstances that had generated the performance-swing. With the help of this method the breweries discovered that the beer-consumption is strongly affected by temperature.

One rule can be for example when somebody is unemployed, young and was employed for a short period formerly then that person's performance will not be adequate or he will notice in a short time. In both cases the fluctuation rate will rise, which will result the restart of the human-resources procurement process, which goes with serious expenses. Recently the realization of the present program is in an initial stage, the program has not been implemented yet. In addition to the implementation of the aforementioned processes and methods, other different data mining methods (as e.g. the claster analyzing) that can effectively be used in the course of staff procurement will also be investigated. [10][11][12][13]

5. CONCLUSIONS

During human-resources procurement the data mining supported selection makes it possible to reduce cost and speed up the process execution. In TESCO department-stores one of the most important human-resources strategy principal is to reduce fluctuation. Applying data mining there is a possibility to carry out this, as based on the processed data it is predictable that the particular applicant how effective and loyal will be.

REFERENCES

[1] LIZÁK, M.: *Személyzetbeszerzés, munkaerő-toborzás in Tothné Sikora Gizella: Humán erőforrások gazdaságtana,* Bíbor Kiadó, 2004. pp. 259-277.

[2] BALOGH, G.: *Emberi erőforrás menedzsment – felsőfokon*, Management Budapest, 2002.

[3] BORGULYA, I., FARKAS, F.: *Emberi erőforrás menedzsment kézikönyv*, KJK-Kerszöv Budapest, 2003.

[4] ADRIAANS, P., ZANTINGE, D.: *Adatbányászat*, Panem Könyvkiadó Kft, 2002.

[5] PARR RUD, O.: *Data Mining Cookbook,* Wiley Publishing Inc., Canada, 2001.

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM

[6] HAN, J., KAMBER, M.: *Adatbányászat koncepciók és technikák*, Panem Könyvkiadó, 2004.

[7] BERRY, M.J.A., LINOFF, G.S.: *Data Mining Techniques,* Wiley Publishing Inc, Indianapolis, Indiana, 2004.

[8] ENSOR, D., STEVENSON, I.: *Oracle tervezés*, Kossuth kiadó, 2000.

[9] GARCIA, H., ULLMAN, M.J.D. WIDOM, J.: *Adatbázisrendszerek megvalósítása*, Panem Könyvkiadó Kft, 2001.

[10] *Sztochasztikus kapcsolatok elemzése* Oktatási segédlet, Miskolci Egyetem Gazdaságtudományi Kar, 2001.

[11] VÁGÓ, ZS.: *Idősorok sztochasztikus modelljei*, Tantárgyi segédlet, 1995.

[12] MICHELBERGER, P., SZEIDL, L.: *Alkalmazott folyamatstatisztika és idősor analízis*, Typotex, 2001.

[13] HUNYADI, L., VITA, L.: *Statisztika közgazdászoknak*, KSH, Budapest, 2002.

DIGITALIZÁLTA: MISKOLCI EGYETEM KÖNYVTÁR, LEVÉLTÁR, MÚZEUM