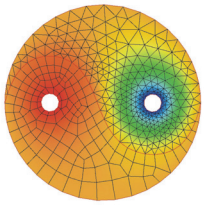
Publishing House
AKAPIT

COMPUTER METHODS IN MATERIALS SCIENCE

Informatyka w Technologii Materiałów

Vol. 9, 2009, No. 3



ASSOCIATION RULES AS AN EXAMPLE OF DATA MINING IN THE ANALYSIS OF COPPER FLASH SMELTING PROCESS – THE METALLURGIST'S POINT OF VIEW

PIOTR JAROSZ, JOLANTA TALAR, JAN KUSIAK

Akademia Górniczo-Hutnicza, al. Mickiewicza 30, 30-059 Kraków
Corresponding Author: kusiak@agh.edu.pl (J. Kusiak)

Abstract

The paper presents an attempt of the exploration of data sets collected in the copper flash smelting process. The goal of the research was to find the association rules occurring between selected output parameters of the process – copper concentration in the slag and lead concentration in blister copper – and input process parameters. The discovered relationships extend the knowledge on the process and can help to improve the models describing the flash smelting process and the process control system algorithms.

Key words: association rules, modelling of metallurgical processes, copper flash smelting

1. INTRODUCTION

Due to a dynamic development of information technologies one can measure and register practically the majority of parameters concerning the analysed industrial process. As a result we have larger and larger, usually exponentially growing amount of data. However, the possibilities of processing such big sets are much smaller than the possibilities of their storing. The problem appears how to use this huge „warehouse" of data to effectively get knowledge, often very interesting and completely unknown if only theoretical analysis of the process was carried out. The tool allowing discovering this hidden knowledge includes the methods of data mining, allowing automatic discovering hidden useful relationships, rules and trends connected with the analysed process in large data sets. In other words, the purpose of data mining is the analysis of data sets for better knowledge and understanding of the processes these data come from. Data mining is a science combining many disciplines: informatics, databases, statistics, artificial intelligence, etc.

One of the methods of data mining includes association rules. This relatively new discipline of knowledge, according to *MIT Technology Review*, will be one of ten technologies that will change the world (Larose, 2005). In a general case one can say that data mining is a process of discovering new important relationships occurring between individual parameters of the process, and they can be found in large quantities of data stored in databases. The above statement means that data mining allows the analysis of databases that (because of their large volume) are difficult to analyse with traditional methods (regression analysis, statistical analysis, etc.). Data mining is also useful in the analysis of such problems where our knowledge on the subject of the analysis is limited.

The main goal of the paper is to present the idea of data mining and method of finding existing asso-

ISSN 1641-8581

ciations in the field of copper flash smelting process. The idea of this method is finding associations in the searched database (i.e. all the relationships and correlations between the parameters of the analysed process, registered in the base). The result consists of the sets of association rules describing the discovered relationships. The discovered rules can sometimes appear unexpected for technologists involved in considered industrial process. On the other hand, discovered rules can be very useful in better understanding of the process and its control.

2. DISCOVERING ASSOCIATIONS

Discovering associations is one of the main areas of data mining. The purpose of this method is finding hidden unknown relationships and links between the data in the analysed large databases. As a result, the association rules (cause-result links) are formulated to describe the discovered relationships. The idea of finding association rules comes from the area of a so-called MBA (Market Basket Analysis), the purpose of which is looking for the patterns of the customers' behaviour, based on the information on the products bought by supermarket customers. A classic example (Larose, 2005) of such analysis is the following rule generated based on the database on supermarket shopping:

„Customers who buy nappies also buy beer".

This rule apparently has no logical sense and is hard to discover with classical, statistical methods. In reality this rule applies to a large population of young fathers shopping in supermarkets. Practical conclusions from the discovered rule can be as follow:

- reduce the price of nappies and increase the price of beer,
- put nappies far away from beer to force the customer to walk across the whole store.

The market basket model makes a certain abstraction, allowing construction of the model of relation *many – to – many* between „products" and „baskets". Association rules based on this model can be defined. In metallurgy an example of a „basket" is classification of grades of copper alloys while the „products" are defined by the chemical elements, which can be treated as impurities of the considered copper alloy. In such a case, an example of a simple association rule can be a statement:

„**if** *the content of phosphorus < 20 ppm and the content of arsenic < 20 ppm*
then *it is a MOOB copper alloy and its electric specific conductivity > 55 MS*".

However, the purpose of data mining is not the repetition of the rules known from theory. It is automatic finding of rules that are unknown from the theory of the analysed process and often overlooked by practitioners.

Association rules are also called *affinity analysis*. These rules take the form:
„if ***antecedent*** then ***conclusion***". A mathematic description of this rule is as follows:

$$A = \{A_1, A_2, ..., A_N\} \Rightarrow B = \{B_1, B_2, ..., B_M\} \quad (1)$$

where: A – attributes vector; B – observation vector.

Intrinsic features of each rule are: a so-called support ***s*** and confidence ***c***. The support ***s*** for a given association rule A ⇒ B is the ratio of the number of observations, fulfilling condition A ⩓ B to the number of all observations. In other words the support of the rule equals the probability of the occurrence of event: A ⩓ B and is expressed with the relationship:

$$s \equiv P(A \cap B) = \frac{number\ of\ events\ A\ and\ B}{total\ number\ of\ all\ events} \quad (2)$$

The confidence of the association rule ***c*** is the ratio of number of observations fulfilling the condition A ⩓ B, to the number of observations fulfilling the condition A. Thus the confidence of the rule is a conditional probability expressed by the relationship:

$$c \equiv P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{number\ of\ events\ A\ and\ B}{number\ of\ events\ A} \quad (3)$$

The values of support and confidence make certain measure used in assessing to which extent the found association rule is interesting. The threshold values for these measures are selected individually and their values depend on the analysed problem. The support and confidence define statistic *strength* of a given rule. The rule is strong, if the values of support and confidence coefficients are greater than certain assumed threshold values; therefore, found rules of high values of both coefficients (support and confidence) are preferred. It is recognized that so-called „strong rules" are those for which ***s*** > 0.20 and ***c*** > 0.70. However, in some cases, when the found rules have much lower values of these parameters and bring new knowledge to the topic and can be useful in the analysis of the problem (analysis by denial).

In the problems of search of the association rules within the Boolean sets, *a priori* algorithms are

commonly applied (Hand et al., 2001). In case of the qualitative data analysis, which we commonly deal with in the analysis of industrial processes, more useful is application of the GRI (Generalized Rule Induction) algorithm (Padhraic & Goodman, 1992).

This algorithm can - for input (antecedent) variables can take both quantitative and qualitative variables, but requires qualitative variables for output (conclusion) variables. The importance of the discovered rule is based on a so-called *J*-measure, defined by the following equation:

$$J = p(A)\left[p(A/B)\ln\frac{p(B/A)}{p(B)} + [1-p(B/A)]\ln\frac{1-p(B/A)}{1-p(B)} \right] \quad (4)$$

where:
p(A) is the probability (or confidence) of the observed value *A*,
p(B) is the probability (or confidence) for values *B*; is to measure the frequency of the occurrence of the conclusion.

As one can see, this measure is a non-linear combination of parameters ***s*** and ***c***. Its property is such that it favours the rules for which the values of the antecedents are more frequent, reflecting better cover in the data set (direct proportionality p(A)). This measure has the higher values also in case when p(B) and p(B/A) reach the limit values (0 or 1). Thus, preferred rules are these, for which the probabilities of the antecedent or confidence are close to their limits.

3. PROBLEM DESCRIPTION

The copper flash smelting process is the worldwide used technology of the smelting of the copper sulphide concentrate elaborated by the Finnish company Outokumpu Oy (Kucharski, 2003). In several plants in the world this process is carried out as the *direct to blister* process, allowing obtaining pure copper (approx. 98.5 % Cu) in one metallurgical aggregate. Such a process is, among others, carried out in Poland. This is a complicated technology, bringing many various problems, both at the stage of chemical, thermodynamic and metallurgical analysis, as well as in the process control. The interest in solving problems occurring during that process leads to apply the search of association rules.

In case of a flash furnace, thousands of parameters are measured. They are of various characters (2593 analogue parameters, 3649 binary parameters and 44 so-called numerator continuous parameters), connected with the chemical composition of charge materials and products of the process, characterizing mass flow in the process, values connected with basic physical parameters (pressure and temperature) in different points of the aggregate, parameters connected with the control, etc. Only in one flash furnace there are several hundreds of the most important parameters measured and registered, then used in the technology analysis and control. Additionally, in case of the analysis of this process, the different frequency of the measurements makes a great problem. Some values, such as, for example, the chemical composition of gas phase, are measured every several seconds. Other values, such as chemical composition of condensed phase, are measured in the period of several hours. Thus the set of the registered values has not only a large volume, reaching 25 MB/24 hrs, but also big differentiation in time series.

Another problem, in case of such big data sets measured in a real industrial object is their quality. Namely, these data are often incomplete due to the failure in the work of sensors and recorders or contain disturbed data. Thus the initial stage of data mining should be the data pre-processing to eliminate disturbances and incomplete data. That problem was described in paper (Stanisławczyk & Kusiak, 2009), where the methods of pre-processing and filtering of the data were presented.

One of crucial problems of the copper concentrate flash smelting, carried out in the considered technology, is a high copper content in the slag. This concentration reaches a relatively wide range from 10 to 16 [% of weight]. Such slag cannot be considered a waste material and undergoes the further processing to regain the copper. High copper concentration in slag is a result of the used technology, which has as one of its goals is obtaining the blister copper of high quality. The quality of the blister copper, produced from polish concentrates, is determined mainly by the adequate low level of lead and arsenic contents in a copper concentrate. In case of smelting concentrates of another composition, these can be concentrations of other metals of higher than copper chemical affinity to the oxygen. To provide low level of the concentrations of these metals in copper, one has to apply a high level of hyperoxidation. This leads to the oxidation of some copper, due to the apparent equalization of copper's thermodynamic reactivity with the reactivity of lead. The copper in the slag would mainly occur in the form of dissolved Cu_2O and in small amount as Cu^{+1} cation – a modifier of silicon-oxygen anions. To

some extent the concentration of copper will be determined by the equilibrium of chemical reaction:

$$[Pb] + (Cu_2O) = 2[Cu] + (PbO) \qquad (5)$$

where: [X] means the component in the metal phase and (Y) - the component in the slag phase.

In thermodynamic analysis it is seen that the concentration of lead in copper can be described in equation:

$$x_{[Pb]} = K^{-1} \frac{x_{(PbO)}\gamma_{(PbO)}}{x_{(Cu_2O)}\gamma_{(Cu_2O)}} \frac{1}{\gamma_{[Pb]}} \qquad (6)$$

when:
x – mole fraction of the component,
γ – activity coefficient of the component, depending on temperature and the composition of the phase,
K – equilibrium constant of the reaction (5), depending on temperature.

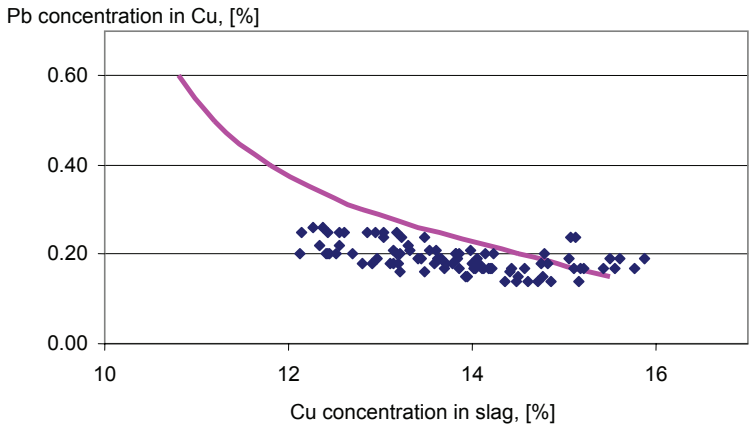


Fig. 1. *The relationship between lead concentration in blister copper and copper concentration in slag (Kucharski, 2003) (continuous line – theoretical curve eq. (6), points – industrial measurements).*

Table 1. *Classification of the values of analysed process parameters.*

Parameter \ Value, [%]	High	Medium	Low
Copper content in a slag	>13	<12.4; 13>	<12.4
Lead content in a blister copper	>0.220	<0.180; 0.220>	<0.180

Figure 1 presents the above theoretical dependence and real values for temperature 1296 °C, selected out of 3751 records measured in an industrial process.

A discrepancy between the measured values and the theoretical curve is seen. These differences increase for low, thus desired in practice, copper concentrations in slag. One of the causes of observed trend can be considering in the theoretical analysis the whole quantity of copper contained in slag to be the dissolved in it Cu_2O. While, in real conditions only the part of copper that was not used as SiO_2 modifier, takes part in the reaction (5). To verify the hypothesis formulated above, an attempt of the analysis of association rules between input and output parameters of the process was undertaken in the present research. The proposed method of seeking association rules can lead to some unidentified so far relationships, allowing the improvement of the control of the process.

The performed calculations were carried out with *Clementine 12* software. The application of the algorithm has to be preceded by the user's definition of the desired minimal levels of support ***s*** and confidence ***c*** of the association rules. It was assumed in the analysis, that ***s*** > 40% and ***c*** > 60 %. Another important feature of the GRI algorithm is also necessary to define the number of association rules that should be found. In this paper the number of rules was limited to five in the considered problem.

4. SOLVING THE PROBLEM

Two process parameters were taken into the consideration for the verification of the assumed hypothesis: concentration of copper in a slag and concentration of lead in a blister copper. According to the technological conditions, these two parameters were classified into the following three groups characterized by technological limits:

The data set of 8792 measurement records was analysed. The process parameters occurring in the search of the association rules were given in the table 2.

The discovered rules relating the analysed parameters and input parameters of the process are listed in the table 3. The applied symbol „&" means the conjunction of parameters. The carried out calculations refer to parameters strongly linked with one another from a technological point of view.

Table 2. *The set of parameters of flash smelting process occurring in the discovered rules (input parameters of the process 1-17, output parameters 18-29).*

No.	PARAMETER	DESCRIPTION	UNIT
1	Concentrate	Total concentrate feed on burners 1-4	[Mg/h]
2	Dusts	Backward dusts	[Mg/h]
3	IOS_P1-5	Total feed of IOS product	[Mg/h]
4	Hyper-oxidation	Hyper-oxidation of the concentrate stream	[Nm^3/Mg]
5	O_2_blow	Mean concentration of oxygen in the blow	[%]
6	Aeration_PK1	Air stream for aeration of solid materials on each burner	[Nm^3/h]
7	Aeration_PK2		[Nm^3/h]
8	Aeration_PK3		[Nm^3/h]
9	Aeration_PK4		[Nm^3/h]
10	Corg	Concentration of basic components in the concentrate	[%]
11	Cu_conc		[%]
12	S_conc		[%]
13	Pb_conc		[%]
14	SiO2_conc		[%]
15	CaO_conc		[%]
16	sub_grains	Parameters characterizing grain composition of the concentrate: fraction of sub- and super-grains	[%]
17	super_grains		[%]
18	T_slag	Mean temperature of the condensed process products in the settler bath	°C
19	T_Cu		°C
20	Cu_slag	Concentration of basic slag components	[%]
21	Fe_slag		[%]
22	Pb_slag		[%]
23	SiO2_slag		[%]
24	CaO_slag		[%]
25	Pb_Cu	Lead concentration in blister copper	[%]
26	CO2_gases	Concentration of basic components of exhaust gases	[%]
27	SO2_gases		[%]
28	O2_gases		[%]
29	NOx_gases		[ppm]

The collected in the table 3 association rules found in the data exploration analysis, in some cases confirm the well known relationships, but in some cases are unexpected according to the "technological knowledge". For example, according to the relationship shown in figure 1, one can expect that the rules associating high copper content in slag and low lead content in copper depend on the same input process parameters (see table 3). However, such relationships were not found as a result of data exploration.

All discovered rules relate the analysed parameters to the process thermodynamics, which is important from technological point of view. Thus these rules appear complementary to each other. This complementariness is seen in the relationships between the analysed output parameters and the chemical composition of the concentrate (concentrations of SiO_2, CaO, Cu, S, Corg., grain composition) as well as the added endothermic components (feed of dusts and IOS product).

In case of the analysis of the lead content in a blister copper, it was also observed the influence of the parameters characterizing the shape of the concentrate stream in the reaction shaft (aeration on burners no. 2 and 3) and oxygen concentration in the blow. No correlation of the analysed parameters and hyper-oxidation was found. That fact is interesting, because that parameter is technologically considered as a basic tool for control of the lead concentration level in blister copper (Talar et al., 2009) on the expert system analysis of considered problem).

The observed supports for the rules of the lead concentration in copper were significantly lower than these for the rules connected with copper concentration in slag, while the confidence levels are opposite. That problem requires further analysis, but initially it can be suggested that it is connected with a different frequency of measurements of these parameters.

The other task of the analysis was a search of the association rules existing between the considered two output process parameters (lead content in a blister copper and a copper content in a slag) and the remaining output process parameters. These relations were searched within the group of the parameters measured with significantly different frequencies (seconds vs. hours, causing the complex structure of the records of the measured data). Linking the parameters of the condensed phases (low frequency of measurements) with the gas phase parameters (high frequency of measurements) would be very useful for the control of the process. This problem is discussed more widely in the paper (Jarosz et al., 2009).

The obtained results of the performed data exploration analysis are shown in table 4. The discovered rules confirm a strong correlation of both analysed parameters. Especially the lead content in copper is strongly linked with the copper concentration

Table 3. *Rules occurring between the analysed output and input process parameters*

Conclusion	Antecedent	Support	Confidence	Conclusion	Antecedent	Support	Confidence
High level of the copper content in a slag	SiO_2_conc > 20.175	43.51	78.32	Low level of the lead content in blister copper	No relationship found	-	-
Medium level of the copper content in a slag	IOS_P1-5 > 5.835 & S_conc < 11.390	65.31	80.33	Medium level of the lead content in a blister copper	C_{org} < 6.675	50.88	100
	IOS_P1-5 > 5.835 & Cu_conc < 30.240	64.25	79.67		Super_grains > 46.150 & Conc < 101.165	43.91	100
	IOS_P1-5 > 5.835	47.57	60.87		C_{org} < 6.675 & Dusts < 10.145	42.19	88.98
Low level of the copper content in a slag	super_grains> 3.050	64.60	84.62	High level of the lead content in a blister copper	IOS_P1-5 > 5.395 & CaO_conc > 5.545	58.00	98.28
					IOS_P1-5 > 5.395 & O_2_blow > 76.965	45.23	91.09
					IOS_P1-5 > 5.395 & Pb_conc < 1.195	45.19	93.21
	super_grains> 3.050 & Dusts > 8.515	59.24	78.79		IOS_P1-5 > 5.395 & Aeration_PK3 > 198.430	44.90	92.12
					IOS_P1-5 > 5.395 & Aeration_PK2 > 197.825	44.82	91.85

in a slag for all levels. From a technological point of view it is confirmed that high copper content in slag should be associated with the slag physical properties (viscosity, melting temperature, surface tension, etc.). The values of these properties are determined mainly by the chemical composition of the phase, which was fully confirmed in the presented results. The influence of the energy parameters of the process on the values of the analysed parameters is also observed. The discovered rules are characterized by the 100% *confidence* and high values of the *support* coefficients. Such high accuracy of the found rules allow, on the basis of continuously measured process parameters (concentration of SO_2, O_2 and NO_x in exhaust gases), to allocate the analysed parameters concerning the condensed phase, to the determined classes. Such classification can be done practically in the on-line mode. From the technological point of view, the found association rules, as well as their high values of the confidence and support indicate, that the combustion process of the concentrate in the reaction shaft occurs in optimal conditions. It also shows that the reaction contribution in the settler bath is low. Moreover, the values of the antecedents of the found rules concerning the concentrations of the exhaust gas phase, although problematic, can be helpful in the control of the process.

The qualitative results obtained using the found association rules show good agreement with the practical knowledge of technologists. However, the quantitative interpretation of these results needs a further analysis of the process experts.

Table 4. *Discovered rules between the process output parameters.*

Conclusion	Antecedent	Support	Confidence	Conclusion	Antecedent	Support	Confidence
High level of the copper content in slag	CaO_slag < 13.46 & T_slag < 1294.5	89.29	100	Low level of the lead content in copper	T_slag < 1293.5 & T_Cu < 1289.0 & NOx_gases < 916.27 & O2_gases < 3.98 & SO2_gases < 15.99 & Cu_slag > 14.11	50.53	100
	Fe_slag > 6.52	86.21	100				
	Pb_slag < 1.93	82.64	100		T_slag < 1293.5 & SO2_gases < 16.115 & NOx_gases < 944.71 & SiO2_slag < 32.96 & Cu_slag > 14.025	46.42	100
	CaO_slag < 13.46	48.18	65.29				
Medium level of the copper content in slag	CaO_slag < 14.83	82.64	100	Medium level of the lead content in blister copper	O2_gases < 2.44 & CO2_gases < 39.49	50.65	100
					O2_gases < 2.44 & CaO_slag < 13.68	49.53	100
					Cu_slag < 12.13	43.90	100
Low level of the copper content in slag	NOx_gases > 2820.485 & O2_gases > 3.57 & CaO_slag < 13.91	97.09	100	High level of the lead content in copper	CaO_slag > 14.65 & CO2_gases > 38.55 & O2_gases < 3.63 & Cu_slag < 12.65 & SO2_gases > 16.41 & T_slag > 1302.5	47.17	100
	Pb_blister > 0.25 & O2_gases > 3.56 & Fe_slag < 5.69	95.24	100		CaO_slag>14.65 & NOx_gases > 897.6 & CO2_gases > 38.50 & SO2_gases < 18.09	47.04	100
	NOx_gases > 2820.48 & CaO_slag < 13.91	90.09	100		CaO_slag > 14.65 & O2_gases < 3.63 & NOx_gases < 1001.12	46.36	100
	Pb_blister > 0.25 & Fe_slag < 5.69	90.09	100		CaO_slag>14.65 & SO2_gases > 16.41 & NOx_gases>897.57 & CO2_gases >38.50	45.70	100

5. CONCLUSIONS

The carried out search of the association rules using the exploration data methods confirms, that such approach can be very useful in the analysis of the flash smelting process. The analysis focused mainly on the examination of the influence of different input parameters on the copper concentration in slag and lead concentration in the blister copper. All found association rules are compatible with the thermodynamical analysis of the considered process.

The discovered association rules are not commonly known, but can be explained theoretically. Therefore, they increase the knowledge about the process itself, and allow the better understanding of occurring phenomena and relations between different process parameters. These rules can improve the model of the flash smelting process and allow increasing the effectiveness of the control algorithms.

ACKNOWLEDGEMENT

The financial support of the MNiSzW, project No 3 T08B 034 30 is acknowledged.

REFERENCES

1. Hand, D., Mannila, H., Padhraic, S., 2001, *Principles of Data Mining*, MIT Press, Cambridge.

2. Jarosz, P., Kusiak, J., Talar, J., 2009, *Analysis of the copper flash smelting process using the data mining, Part II – Association Rules*, Rudy i Metale Nieżelazne, 7 (in press)
3. Kucharski, M., 2003, *Pirometalurgia miedzi*, AGH, Krakow (in Polish).
4. Larose, D.T., 2005, *Discovering knowledge in data: an introduction to data mining*, J. Wiley & Sons, Hoboken.
5. Padhraic, S., Goodman, R.M., 1992, An informative theoretic approach to rule induction from database, *IEEE Transaction on Knowledge and Data Engineering*, 4, 4, 301-316.
6. Stanisławczyk, A., Kusiak, J., 2009, Pre-processing of the industrial data for data mining and modelling – application to the copper flash smelting process, *Computer Methods in Materials Science*, 9, 3, 369-373.
7. Talar, J., Jarosz, P., Kusiak, J., 2009, Expert system application in modelling and controlling the copper flash smelting process, *Computer Methods in Materials Science*, 9, 3, 379-391.

REGUŁY ASOCJACYJNE JAKO PRZYKŁAD EKSPLORACJI DANYCH W ANALIZIE ZAWIESINOWEGO PROCESU WYTOPU MIEDZI – SPOJRZENIE METALURGA

Streszczenie

W ramach niniejszej pracy przeprowadzono eksplorację danych, pochodzących z procesu otrzymywania miedzi w piecu zawiesinowym. Celem pracy było odszukanie reguł asocjacyjnych występujących pomiędzy wybranymi parametrami wyjściowymi procesu – stężeniem miedzi w żużlu i stężeniem ołowiu w miedzi – a parametrami wejściowymi. Znalezione zależności poszerzają wiedzę o procesie i mogą być pomocne w zakresie poprawy modeli opisujących proces zawiesinowy oraz algorytmów sterowania procesem.

Received: April 21, 2009
Received in a revised form: May 25, 2009
Accepted: June 15, 2009