http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (19): 5264-5269, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5264.5269

Data Mining System for Triazophos Synthesis Process

[1]Zhang Quanling, [1]Gu Yong, [1]Xu Weihua, [2]Hong Yanping and [2]Zhao Lujun
[1]Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, 310027, P.R. China
[2]Zhejiang Supcon Software Co., Ltd., Hangzhou, 310053, China

Abstract: Data Mining (DM) is the process of extracting desirable knowledge or patterns from existing databases or dataware house for specific purposes. Much research has been done focusing on its application to retail sales, but it is seldom applied in process industry. Based on the characteristic of process industry and association rules with multiple minimum supports, a new algorithm of data mining which can be applied in process industry is proposed and a total solution for data mining system in triazophos synthesis process is suggested. The solution is composed of DCS-based control system, real-time database, data preprocessing, data mining and visualization of data-mining results. Its application in triazophos synthesis process brings about more economic benefits for enterprise and is highly appraised by science and technology department of Zhejiang province, PRC.

Key words: Triazophos synthesis process, data integration, data mining, association rules

INTRODUCTION

The large number of data and their corresponding information is a huge asset for an enterprise. However, there is a common problem existing in all enterprises: in such a huge database, the valuable information takes up a small amount. That is to say, a powerful data mining tool is needed to extract useful information. Without such a powerful tool, the large number of information in the database would be "data tomb" which could not be accessed effectively. Meanwhile, without properly extracting information from the large database, decision-makers could only make decisions based on their subjective judgment instead of objective study. To retrieve useful information from the database, an expert system was proposed in that customers and expertise input the knowledge into the database. In this sense, the expert system is subject to constant errors and deviations and meanwhile, it is expensive and time-consuming.

Tools based on data-mining system are to analyze the large number of data and explore the intrinsic important data patterns. They are conducive to decision-making, knowledge database establishment and scientific research. With such tools, the huge gap between data and information can be bridged up and the data tomb can be turned into golden stones. Therefore, it comes to be a priority on how to explore the operation pattern based on the huge database and provide valuable information to the decision making process and finally bring great economic benefits to the enterprises. Data mining is one of the solutions. It is an in-depth data mining tool and it can be described as follows: it is an advanced and efficient way to explore and analyze a large number of data, to reveal some latent, unlearnt or testified rules and to make them patterned.

Currently, data mining has become an efficient way to analyze database or dataware house (Bi and Zhang, 2005; Lee *et al*., 2005; Tsai and Chen, 2004; Huang and Yin, 2003; Lou and Baokang, 2001; Hong and Wu, 2011, Hamid *et al*., 2011). It is aimed to extract desirable knowledge out of existing database and demonstrate to users in a friendly way. Simply put, data mining is to extract the latent, unlearnt, constructive knowledge for decision-making process from a large number of incomplete, noisy, vague and random data at work. Data mining is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematic algorithms and statistics. Given the enormous size of databases, data mining is a technology for Knowledge Discovery in Databases (KDD). This technology provides different methodologies for decision-making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning and innovation (Liao, 2003).

MINING ASSOCIATION RULES WITH MULTIPLE MINIMUM SUPPORTS

Mining association rules: Mining association rules was first introduced by (IBM Almaden Research center) Agrawal *et al*. (1993). Mining association rules is a technique that retrieves all possible patterns from very

Corresponding Author: Zhang Quanling, Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, 310027, China

large databases that match some pre-defined conditions, by which the results obtained can be used for future predictions. An example of an association rule is "20% customers purchase both bread and milk, 70% customers who purchase bread also purchase milk." This association rule can be expressed in the form:

$$\text{bread} \rightarrow \text{milk } [S = 20\%, C = 70\%]$$

where, S is support, C is confidence.

Generally, the basic principle of mining association rules can be shown as below (Lim and Lee, 2010).

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of literals (called items). Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in T, if $X \subseteq T$.

Generally, an association rule is an implication of the form:

$$X \rightarrow Y$$

where, $X \subset I$, $Y \subset I$ and $X \cap Y = \Phi$. X is known as the antecedent and Y is known as the consequent. The rule $X \rightarrow Y$. holds in the transaction set D with confidence C if (C*100)% of transactions in D that contain X also contain Y. The rule $X \rightarrow Y$ has support S in the transaction set D if (100*S)% of transactions in D contain $X \rightarrow Y$.

The support value is defined as the probability that the items of X and Y are all present in the database. The confidence is the probability that if a record contains the items of X, that same record contains also the items of Y. The support and confidence of the itemsets (X) and (Y) for a rule from a database of T records can be calculated as follows:

$S(X \rightarrow Y)$ = probability $(X \cup Y)$ = {records with X and Y}/{all records}

$C(X \rightarrow Y)$ = probability $(Y|X)$ = {records with X and Y}/{records with Y}.

For both S and C we define a lower threshold, min_sup and min_conf. Given a set of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence greater than min_sup and min_conf respectively. Every rule that has support and confidence superior to min_sup and min_conf is said to be a "strong" rule.

An association mining algorithm works in two steps:

- Generate all large itemsets that satisfy min_sup
- Generate all association rules that satisfy min_conf using the large itemsets. An itemset is simply a set of items. A large itemset is an itemset that has transaction support above min_sup

The key element that makes association rule mining practical is the min_sup. It is used to prune the search space and to limit the number of rules generated. However, using only a single min_sup implicitly assumes that all items in the data are of the same nature (to be explained below) and/or have similar frequencies in the database. This is often not the case in real-life applications. In many applications, some items appear very frequently in the data, while others rarely appear. If the frequencies of items vary a great deal, we will encounter two problems (Liu *et al*., 1999):

- If min_sup is set too high, we will not find those rules that involve infrequent items or rare items in the data
- In order to find rules that involve both frequent and rare items, we have to set min_sup very low. However, this may cause combinatorial explosion, producing too many rules, because those frequent items will be associated with one another in all possible ways and many of them are meaningless

Mining association rules with multiple minimum supports (Liu *et al*., 1999): In our extended model, the definition of association rules remains the same. The definition of minimum support is changed. In the new model, the minimum support of a rule is expressed in terms of Minimum Item Supports (MIS) of the items that appear in the rule. That is, each item in the database can have a minimum item support specified by the user. By providing different MIS values for different items, the user effectively expresses different support requirements for different rules.

Let MIS(i) denote the MIS value of item i. The minimum support of a rule R is the lowest MIS value among the items in the rule. That is, a rule R, a1, a2, …, $ak \rightarrow ak+1, \ldots,$ ar where aj I, satisfies its minimum support if the rule's actual support in the data is greater than or equal to:

min(MIS(a1), MIS(a2), …, MIS(ar)).

Minimum item supports thus enable us to achieve the goal of having higher minimum supports for rules that only involve frequent items and having lower minimum supports for rules that involve less frequent items.

For Example:

Consider the following items in a database, bread, shoes, clothes. The user-specified MIS values are as follows:

MIS(bread) = 2% MIS(shoes) = 0.1% MIS(clothes) = 0.2%

The following rule doesn't satisfy its minimum support:

clothes→bread [sup = 0.15%, conf = 70%]

because min(MIS(bread), MIS(clothes)) = 0.2%. The following rule satisfies its minimum support:

clothes→shoes [sup = 0.15%, conf = 70%]

because min (MIS(clothes), MIS(shoes)) = 0.1%

While a single min_sup is inadequate for applications, we also realize that there are deficiencies with min_conf of the existing model.

DATA MINING SYSTEM FOR TRIAZOPHOS SYNTHESIS PROCESS

Because the multiple minimum supports are exclusively used to analyze binary data, while the process parameter and product data of trizophos synthesis process (Zheng *et al*., 2001) are decimal data, therefore, the multiple minimum supports can not be applied. This study proposes an improved algorithm of mining association rules with multiple minimum supports. It can be described as follows:

Target of data-mining in triazophos synthesis process: The aim of data mining in triazophos synthesis process is to find out the basic process parameter based on resourceful real-time data and propose optimized recipes, process parameters to improve the productivity and triazophos contents and finally bring more economic benefits for enterprises.

Data mining algorithms in triazophos synthesis process: Given:

- Triazophos synthesis process has n parameters, A1, A2, …, An
- S is product yield
- C is triazophos content

Then the data mining algorithm of triazophos synthesis process can be described as follows:

Define the initial operation parameter and its predictive product yield and triazophos volume: In order to realize the data mining purpose, let the initial operation parameter Init_Ai (i = 1, 2, …, n) be the mean of production history database. Because the purpose is to obtain the highest triazophos contents, the predictive product yield Init_S and triazphos containts Init_S are defined as the mean of history database.

Transform the production history database into binary database.

Transform the production history database (DB_OLD) containing digital parameter and its relevant product yield, content data into binary database (DB_NEW):

- Transformation of database structure

Transform the triazophos synthesis process (DB_OLD) database structure (that is, every operation condition A1, A2, …An), triazophos product yield (S) and triazophos content (C) into a new database structure (DB_NEW) to be mined, including A1_Increase, A1_decrease, A2_Increase, A2_decrease, …An_Increase, An_decrease, S_Increase, S_decrease, C_Increase and C_decrease.

Transform the original real-time database into to-be-mined database: Each record of the original real-time database (DB_OLD) is compared with the initial operation parameter Init_Ai (i = 1, 2, …n), expectational product yield (Init_S) and trizophos content (Init_C) and thus a corresponding new record in the to-be-mined database is generated. The transformation procedure can be illustrated as follows:

Let production data in real-time DB_OLD be A1, A2, …, An, S, C, then the corresponding data in DB_NEW will be A1_Increase, A1_decrease, A2_Increase, A2_decrease, …, An_Increase, An_decrease, S_Increase, S_decrease, C_Increase and C_decrease. The transformation is shown as follows:

- If A1<Init_A1 then A1_Increase = 0, A1_decrease = 1 else A1_Increase = 1, A1_decrease = 0
- If A2<Init_A2 then A2_Increase = 0, A2_decrease = 1 else A2_Increase = 1, A2_decrease = 0

  … …
- If An<Init_An then An_Increase = 0, An_decrease = 1 else An_Increase = 1, An_decrease = 0
- If S<Init_S then S_Increase = 0, S_decrease = 1 else S_Increase = 1,S_decrease = 0
- If C<Init_C then C_Increase = 0,C_decrease = 1 else C_Increase = 1, C_decrease = 0

Association rules with multiple minimum supports algorithms for DB_NEW data mining: Let S_Increase be target item, let "A1_Increase, A1_decrease, A2_Increase, A2_decrease, …, An_Increase and An_decrease" be relational items, conventional association rules with multiple minimum supports algorithms are adopted for data mining and strong correlational rules are obtained. Given there are k rules that satisfy operation conditions, let Rj (j = 1, 2, …, k), then the Rj can be described as:

- Rj: S_Increase→relational items (Confidence, minimum support)

Rules interpretation: The results of data mining indicate:

- Rj: S_Increase→relational items (Confidence, minimum support)

An example is used to illustrate this rule.

Given:

- Let R1: S_Increase _ A2_Increase, A4_Decrease (70%, 51%)
- Let total number of records in DB_NEW be NT
- Let total number of records that satisfy the following two conditions be NA, one condition being A2>Init_2, the other condition being A4<Init_4
- Let total number of records that satisfy the following three conditions be ND, one condition being being A2>Init_2, the second being A4<Init_4, the third being production yield S>Init_S

Then ND/NT = 51%, ND/NA*100% = 70%

That is to say, R1 denotes 70% of all records that satisfy A2>Init_2 and A4<Init_4 also satisfy S>Init_S.

Optimized operation conditions for rule Ri: Even though after data mining process, k rules have been obtained that satisfy minimum supports and minimum confidence, these rules, however, are of qualitative function rather than of quantitative value. This is not the ultimatum purpose of triazophos synthesis process. Then how to obtain the quantitative value poses a question which association rule with minimum support can not solve.

In this algorithm, to find out the optimized operation conditions of each rule, we trace back to the real-time database, in this sense, if we want to search for R1: S_Increase→A2_Increase, A4_Decrease (70, 51%), then it can be operated as follows:

- Search DB_NEW for all records that satisfy A2>Init_2, A4<Init_4 and S>Init_S, let the total number be ND and record them as RN (i), i = 1, 2, …, ND
- Search DB_OLD for ND records corresponding with RN(i) and record them as RO(i), i = 1, 2, …, ND
- Locate the record of highest S among RO(i), i = 1,2,…,ND and record its optimized operation conditions as OPT_A1,OPT_A2, …, OPT_An
- Then OPT_A1,OPT_A2, …, OPT_An are the corresponding optimized operation conditions of R1.

Rule application: The data mining rules apply to control process on the condition of careful consideration of process characteristics by process engineers.

Data mining technical framework of Triazophos synthesis process: Based on DCS (Distributed Control System), the data mining technical framework is aimed to establish history database on the basis of real-time database(Ye, 2005), optimize production process, operation condition and recipe though association rule with multiple support. The technical framework of data mining system is illustrated in Fig. 1.

Data integration for triazophos synthesis process: Generally, Triazophos Synthesis Process data can be divided into two categories, online data and offline data.

- **Online measurements data:** Generally, online measurements data is stored in Distributed Control System (DCS) or real-time database. Data mining system has interface for typical real-time database which can integrate data from real-time database. Furthermore, standard OPC interface is provided which enables any OPC Server to communicate with the data mining system. This means no new drivers are needed for any source that provides OPC Server support

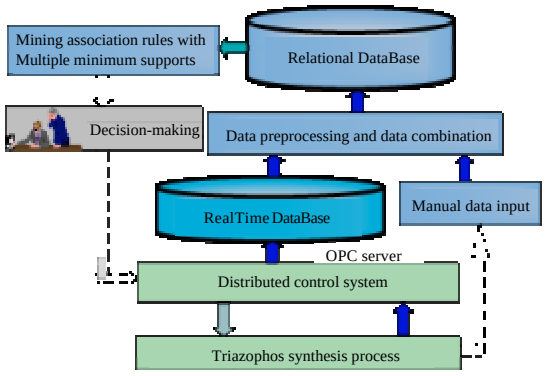

Fig. 1: Technical framework for data mining system

- **Offline measurement data:** Besides online measurement data, many other data such as product quality data that can only be obtained via assay are input into system manually and subsequently integrated with real-time data. This system also provides an interface for manual data input
- After data preprocessing, the real time-data and manually input data are integrated into the relational database for future use
- **Data mining process:** Before data mining operation, relevant parameter should be set up and its configuration is shown in Fig. 2. When the configuration is finished, data mining process can be started.

If data mining operation starts to generate rules (more than one rule), then graphs with various data mining results can be formed as illustrated in Fig. 3. In this graph, click the buttons of "Rule Pie", "Rule Arrows", "Optimization Proposal", "Optimized Operation Condition," we can get different visualizations of data mining results.

Data mining application: Considering parameter confidentiality requirements, the production data and recipe in triazophos synthesis process have not been mentioned. The real-life application of this system can be summarized as follows (abridged from customer report): The operation indicates that after data mining recipe and

Fig. 2: Configuration of data mining system

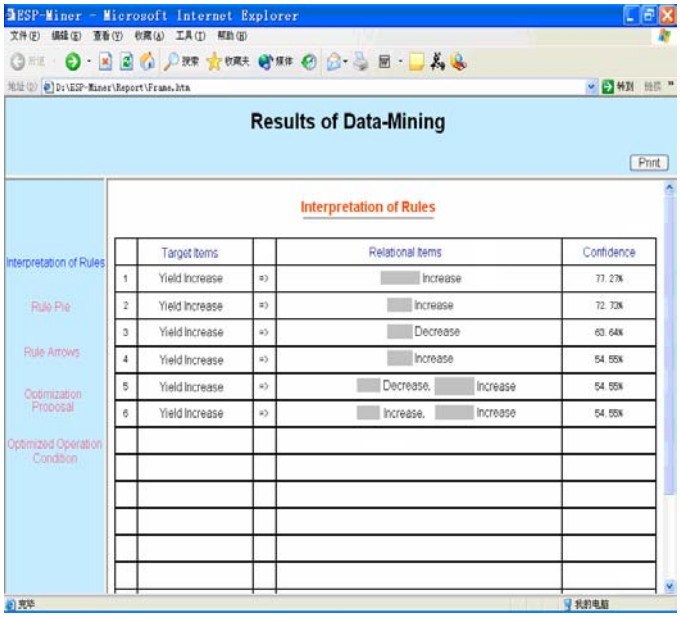

Fig. 3: Visualization for results of data-mining system

optimized parameter are adopted, the product yield of triazophos is increased by 1.5% on average and the average triazophos content increases from 80-83%.

CONCLUSION

The data-mining system for triazophos synthesis process is composed of DCS-based control system, real-time database, data preprocessing, data mining and visualization of data-mining results. Based on algorithm of association rules with multiple minimum supports, a new algorithm of data-mining is proposed. Using this algorithm, the data-mining software ESP-Miner for process industry is developed. The ESP-Miner enables enterprises to find out the relationship between the product yield, triazophos content and recipe and process control condition. Furthermore, the modification trends of process parameter and control condition are proposed. The application brings about more economic benefits for enterprise and is highly appraised by science and technology department of Zhejiang province, PRC.

ACKNOWLEDGMENT

The authors would like to thank for the support by the National High Technology Research and Development Program of China (863 Program 2012AA040307, 2009AA043204) and the national science and technology support program (No. 2012BAF05B01).

REFERENCES

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 25-28, 1993, Washington, DC., USA., pp: 207-216.

Bi, J. and Q. Zhang, 2005. Data mining association rule algorithms survey. Eng. Sci., 4: 88-94.

Hamid, R.Q., N. Mahdi and M.B. Behrouz, 2011. Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. Expert Syst. Appl., 1: 288-298.

Hong, T.P. and C.W. Wu, 2011. Mining rules from an incomplete dataset with a high missing rate. Expert Syst. Appl., 4: 3931-3936.

Huang, J. and Z. Yin, 2003. Improvement of apriori algorithm for mining association rules. J. UEST Chin., 1: 76-79.

Lee, Y.C., T.P. Hong and W.Y. Lin, 2005. Mining association rules with multiple minimum supports using maximum constraints. Int. J. Approximate Reasoning. 40: 44-45.

Liao, S.H., 2003. Knowledge management technologies and applications-literature review from 1995 to 2002. Expert Syst. Appl., 25: 155-164.

Lim, A.H.L. and C.S. Lee, 2010. Processing online analytics with classification and association rule mining. Knowledge-Based Syst., 23: 248-255.

Liu, B., W. Hsu and Y. Ma, 1999. Mining association rules with multiple minimum supports. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, CA, USA, pp: 337-341.

Lou, X.H. and D.G. Baokang, 2001. An association rule mining algorithm based 0n multiple supports. Comput. Eng., 27: 102-103.

Tsai, P.S.M. and C.M. Chen, 2004. Mining interesting association rules from customer databases and transaction databases. Inform. Syst. J., 29: 685-696.

Ye, J.W., 2005. Key technology and system architecture of large-scale real-time database. Master Thesis, Zhejiang University, China.

Zheng, Z.M., L.X. Li, H.B. Yu, X.H. Zhang, H. Lin and Z.J. Wang, 2001. The new synthesize method of triazophos. Agrochemicals, 40: 14-14, 21.