Mining of EDI Data for Performance Measurement of a Supply Chain

Trung T. Pham
DIcentral Corporation
tpham@DIcentral.com

***Abstract*. This paper presents an application of data mining using EDI data to measure the performance of a supply chain. This performance is measured in terms of the turn-around time for a business transaction. The data mining problem focuses on a large set of EDI data in which the formulation to calculate the turn-around time is well defined and the parameters involved are clearly identified. One issue associated with mining EDI data is the computational load in parsing the data is also discussed. Results of some EDI data sets are summarized and presented as demonstration for data visualization.**

I. Introduction

Data mining [1-5] is the process of extracting knowledge from a large set of raw data. It is common practice for the data to be stored in a database for efficient retrieval. The extracted knowledge is often used for post-operation analysis, supporting a decision making process, or as input into another application.

The process of data mining traditionally involves several subtasks, e.g., exploratory data analysis [6-7], descriptive modeling [8-9], predictive modeling [10-11], etc. These subtasks are used to address specific needs and requirements that the owner of the data might have. Sometimes the data are used to discover unexpected knowledge outside the pre-defined needs and requirements.

In an inter-organizational business-to-business (B2B) process [12-14], business transactions are flowing from one organization to another. These transactions represent commercial activities such as trading, planning and scheduling, monetary wiring, authorizing, etc. These transactions are often stored in a database connected to an Enterprise Resource Planning (ERP) system [15-16]. In addition to the direct usage of these transactional data for commercial trading, this database can also be used to provide information that might help improve the productivity/performance of a business process.

The use of electronic commerce has increased the flow of these transactions to a faster rate due to an increase in efficiency in data handling. In this setting, an organization buying products and services from a supply chain (a large number of suppliers) must rely on the fast pace of transactions to plan its ordering activities. To minimize the inventory cost, an organization will utilize on the performance of its supply chain to plan its "just-in-time" inventory system.

In order to achieve a certain performance, quantitative measure must be defined, tracked, and controlled. In the management of supply chain, the response time to a purchase order is used to reflect the performance of a supplier. This quantitative measurement is tracked against a set of dependent variables to help manage and control the performance to a satisfactory level.

This paper presents the use of mining EDI data to measure the performance of a supply chain. In this setting, an organization managing a large supply chain can record the turn-around time of a purchase order as a parameter reflecting the performance of each supplier. This is the time elapsed between the start of a purchase order and the receipt of an invoice. The performance of a supply chain is assumed dependent on a set of parameters, some directly retrievable from the data on an invoice, some from a database of suppliers.

One side issue in mining EDI data is the parsing of a data set compactly packed in a particular format [17-18]. This phenomenon is the result of EDI data being packed into a compact size to minimize the data transmission cost [19-20]. In order to retrieve a data element, an EDI data message must first be parsed into several segments, each representing a specific type of data and containing a number of data elements. The parsing of EDI data requires some computational effort that can accumulate into a sizeable burden when a large number of EDI data messages is involved.

II. Mining EDI Data To Measure Performance

This section summarizes the basic tasks of the data mining process for measurement of the performance of a supply chain.

An organization trying to optimize its inventory system must rely on the response time of its suppliers. This response time is critical for the planning of a just-in-time inventory system to minimize the inventory cost and to maximize uninterrupted operation period. These factors play major roles in allowing the organization to retain competitive advantages over its competitors.

A. Assumptions

It is assumed that the suppliers share the same EDI environment, e.g., the same service provider, the same computer capability, and the same business rules. This assumption is a fair factor to help eliminating some basic variations that might affect the analysis of a large supply chain. For a large supply chain, the suppliers can be categorized into different groups, each sharing the same EDI environment so that the analysis can focus on fewer (and perhaps, a manageable number of) variables.

B. Observable Output

The response time of a supplier is measured as the time duration between the time a purchase order is placed and the time the corresponding invoice is received. Through the use of electronic data interchange, the period when these documents are in transit is practically reduced to fraction of a second, leaving the response time as a true indicator of how efficient a supplier is set up.

Alternatively, other parameters can be used as the measurement of the response time. One can use the duration between the time a purchase order is placed and the time the shipping notice is received. This parameter measures the total cycle of the ordering process, leaving out the details of what might happen in each step in between. Another parameter that one can use as the measurement of the response time is the duration between the time a purchase order is placed and the time of shipment arrival. Again, this parameter includes the performance of the carrier, which does not accurately reflect just the performance of a supplier.

In the scope of this paper, the duration between the time a purchase order is placed and the time the corresponding invoice is sent is used. A typical invoice document normally contains both the purchase order timestamp and the invoice timestamp. Each timestamp routinely details the date, month, year, hour, minute, and second. These data allow the calculation of the response time through a simple subtraction algorithm of two different timestamps.

Since the calculated response time is deterministic and based on the time set by computer systems, this output measurement can be considered unbiased and not contaminated by any interference. Occasionally, a supplier might stamp an invoice with time from a different time zone, or unethically change the time on its computer system to manipulate the data to his advantage. To address these issues, an organization can override the invoice timestamp with the time the invoice is received (instead of the time the invoice is sent). This practice is acceptable in the use of the data in an organization's ERP system to measure the response time. However, the timestamp at the moment an invoice is sent is still widely accepted as legally binding for other purposes.

C. Dependent Variables

There are two sets of dependent variables in the process of measuring the performance of a supplier. One set of variables can be retrieved directly from the invoice, and another set can only be retrieved from a database containing supplier information. These dependent variables will be use to help the analysis process that tries to extract an explanation to why a supplier or a group of suppliers is performing at certain rate. These variables are hypothesized, with reasons presented below, to play major roles in affecting the performance.

From an invoice, the following variables are retrieved: the total amount of an invoice, the number of line items in an invoice, and the number of individual items in an invoice.

In general practice, it is common sense to expect that a supplier would systematically process purchase orders with the first-in-first-out basis. However, it is widely practiced for a supplier to process a purchase order with the largest monetary total amount first because that purchase order generates more income.

A supplier might prioritize the order of processing purchase order according to the number of line items involved. This scenario is normally encountered with large suppliers who want to be more efficient in packing the delivery shipment: the purchase orders with fewer line items will be processed by a different group for efficiency while the ones with more line items will be handled separately by a more experienced group. In addition, however a purchase order is handled, it is always expected to take more time for the orders with more line items. Similarly, the total number of individual items in a purchase order can affect the rate it is processed.

From a database containing supplier information, the following variables are retrieved: the number of buyers a supplier must deal with, the geographical region a supplier is in, and the number of products a supplier is selling.

The number of buyers a supplier must deal with is a major factor in determining how fast the response time can be. If a supplier is an exclusive supplier, i.e., it deals with only one buyer, the expected performance is drastically different from the case when a supplier must deal with a large number of buyers. Another factor in analyzing the performance of a supplier is the number of product items it is offering: the more items there are in its catalog, the slower a supplier might be in and invoice preparation process that includes determining the availability, the storage location, the freight companies, the packing and labeling, etc. The geographical regions that a supplier might be in also play a role in determining the freight companies and in reflecting the style of working.

D. Classification Definitions

The performance of a supplier can be categorized into three groups: (i) preferred supplier (fast response), (ii) normal supplier (medium or acceptable response), and (iii) non-performing suppliers (slow response).

Since the Electronic Data Interchange process is implemented solely for the purpose of increasing the rate of the data flow, it is reasonable to expect that a purchase order be attended to within 24 hours. This expectation is translated into a preference and the suppliers who satisfy this preference is normally given a preferred rating that plays a major role in determining who should be on the approved supplier list in the long run.

A normal supplier is the one who consistently responds within two business days in the evaluating period. This number is the average response time, with a small variance. This variance number shows the acceptable range outside this normal level. Normally, a variance value of less than 1 is used.

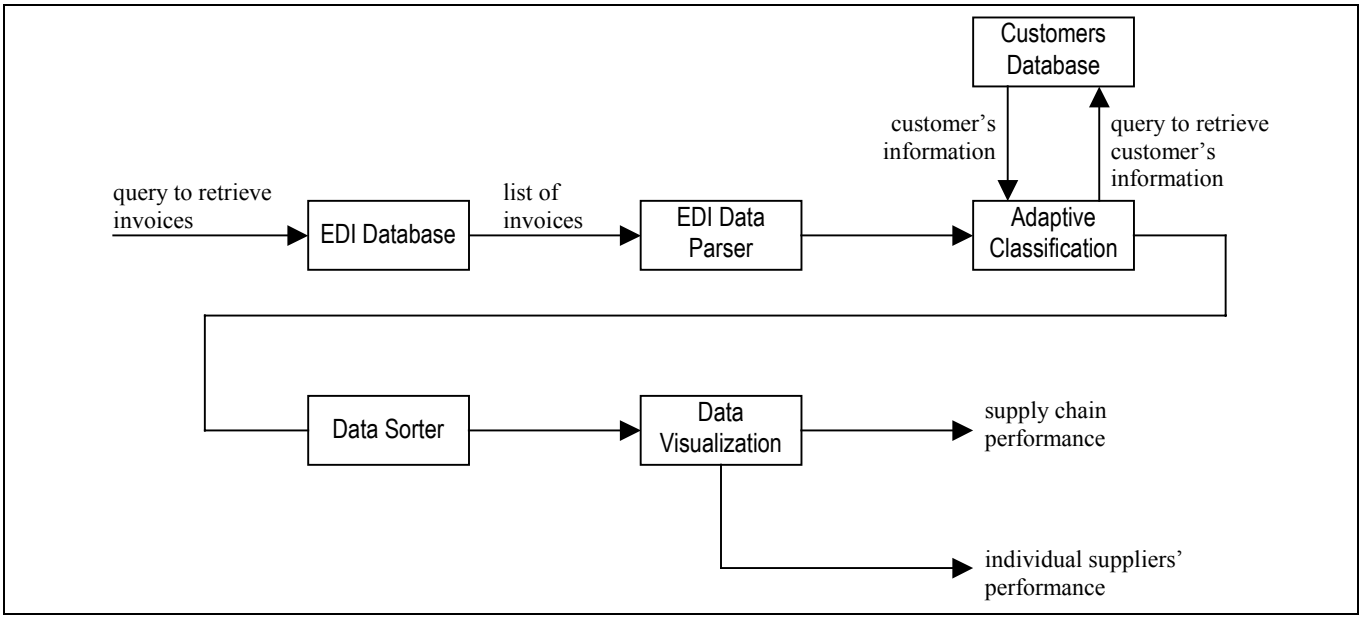

Figure 1. Data flow in the mining process of evaluating supply chain performance.

A non-performing supplier is the one who takes more than two business days to respond to a purchase order. Since the response time is already outside the acceptable level, the variance range is not crucial in determining which supplier is not performing. However, the variance range might be helpful in analyzing the individual failure for explanation that might help in managing a large supply chain.

III. EDI Data Computational Requirements

The complete process of mining EDI data to evaluate the supply chain performance consists of the following four basic steps: (i) query appropriate data from a large database, (ii) parse the retrieved data into data elements, (iii) classify parsed data into different categories, and (iv) prepare graphical reports for analysis. Figure 1 depicts the data flow in this process.

A. Data Query Procedure

A database containing EDI transactional data is queried for the purpose of retrieving useful information to determine the performance of a supply chain. In the context of this paper, only invoice transactions are retrieved. Furthermore, the data are retrieved for a certain time period of evaluation. Two query criteria are therefore used: transaction type (invoice), and transaction date (invoice date).

Since it was common practice to store raw EDI data in a database, the query must look for a string containing certain signatures representing the search criteria. An EDI message is a string of several characters compactly packed together. The transaction type is encoded in a segment with the segment identification "ST". For an invoice, the first element of this segment is a string constant "810." As the result, the search query will look at the raw EDI data string and search for a sub-string

```
ST*810
```

embedded in it. Notice the use of an asterisk (*) in this search string: it is the standard delimiter used to separate the data elements and the segment identification.

A second criterion to be used in the query is the invoice date. An EDI segment containing the invoice date has the identification "DTM" and the qualifier "003" in its first element. The second element is the invoice date in the format of 8 characters, with the first four characters representing the year, the next two characters representing the month, and the last two characters representing the date. The search of an invoice in a particular date must contain the following sub-string

```
DTM*003*YYYYMMDD
```

where the string YYYY represents the year, the string MM represents the month, and the string DD represents the date.

A typical query command looking for all EDI messages that are invoices in a particular date can be written in an SQL-like format:

```
SELECT data FROM EDItable
WHERE
   data = '*ST\*810*' AND
   data = '*DTM\*003\*YYYYMM*'
```

B. EDI Data Parsing Procedure

Once an EDI message is retrieved from a database, it must be parsed into a structure containing several segments, each segment containing individual data elements. The parsing of an EDI message is particularly simple because of a well-defined convention separating data segments and data elements: a pre-defined symbol, the circumflex accent (^), is used as a delimiter separating the segments, and another pre-defined symbol, the

asterisk (*), is used as a delimiter separating the segment identification and data elements. The parsing algorithm consists of two separations, one to separate the segments, and another one nested in a loop to separate the segment identification and the data elements.

For most of the time, an invoice will contain the purchase order date in the segment DTM with the first data element serving as a qualifier containing the constant string "004." However, sometimes a supplier's computing capability is not sophisticated to handle that reference, the buyer must query back into its database to cross-link the invoice with the original purchase order to retrieve this purchase order date. In this case, the invoice must contain the original purchase order number to allow this cross-linking. The parsing procedure will be slower due to an additional query into the buyer's database to retrieve more information.

Since every EDI data message must be parsed before its data are used, the computational load of this mining procedure is a major issue in dealing with a large database containing many EDI data messages. An alternative approach to reduce the run time of such a query is to parse data as they come in. This approach will not reduce the total parsing time but will significantly reduce the mining time. One hidden cost to this approach is the complexity of a table designed for storing the already parsed data. The number of variables must be pre-set at a large default number to cover all possible cases. This practice might slow down the search procedure when dealing with a more complex table.

C. Data Classification

Invoices during a particular review period are pulled out from an EDI database. These invoices come from many suppliers and will be classified on the averaging basis. As the result, the following adaptive algorithm is implemented to categorize supplier's performance.

For an individual supplier, the average response time is updated every time an invoice that it sent out is encountered. The update average time is calculated as follows:

$$\mu_{n+1} = \frac{n}{n+1}\mu_n + \frac{1}{n+1}x_{n+1},$$

where x_{n+1} is the $(n+1)^{\text{th}}$ response time point encountered, and μ_n the average response time based on n points. Similarly, the variance is updated as follows:

$$\sigma_{n+1}^2 = \frac{n}{n+1}\sigma_n^2 + \frac{1}{n+1}(x_{n+1} - \mu)^2.$$

The second formula is the exact iterative formula where the average μ is assumed to be calculated in the first pass of a two-pass process. However, in a real-time one-pass process, it is common practice to use the estimated average at that particular time and the resulting error is bounded within some acceptable limits. These limits are determined by the statistical model and the number of sample data used [21].

These two parameters are used to determine the classification of a supplier. The algorithm for classification is as follows:

(i) if $\mu \le 1$, supplier is in preferred list
(ii) if $1 < \mu \le 2$ and $\sigma^2 \le 1$, supplier is in normal list
(iii) if $1 < \mu \le 2$ and $\sigma^2 > 1$, supplier is in "concerned" list
(iv) if $\mu > 2$, supplier needs to improve on performance

D. EDI Data Organization and Report Preparation

The data organization process consists of grouping data as a whole to analyze a supply chain, and grouping data according to individual supplier to analyze each supplier for performance improvement (if needed).

For a report preparation, data must be presented in some graphical representation allowing a quick analysis that can form a hypothesis of a relation linking the performance level with the dependent variables.

IV. Simulation Results

This section outlines the numerical results in extracting performance information from a database storing EDI data. The results will be organized into two categories: the performance of an overall supply chain, and the performance on individual suppliers.

A. Overall Supply Chain Performance

The performance of an overall supply chain reflects a brief summary of all suppliers. Typically, a histogram listing the number of companies responded in a particular timeframe (Fig. 2). This histogram should show a heavy concentration of suppliers around an acceptable level of 1 day. There will be a small number of suppliers on the borderline of 2 days, and a very small number of suppliers outside the acceptable range.

A detail report of the supply chain will list out individual supplier names according to each category (Fig. 3). This report will be useful for a buyer to single out non-performing suppliers to either schedule meeting for enforcement or to drop them off the approved list. This report is often printed in colors to help human easily categorize the suppliers according to each pre-assigned color.

The performance of an overall supply chain will also be plotted in a three-dimensional chart against any of the dependent variables listed earlier: total amount of an invoice, the number of line items in an invoice, the number of individual items in an invoice, the number of buyers a supplier must deal with, the geographical region a supplier is in, and the number of products a supplier is selling. These plots will help a supply chain manager analyze the performance in detail and form a hypothesis about a relation between the performance and these variables. Figure 4 depicts such graphical results.

B. Individual Supplier's Performance

Individual supplier's performance is also needed to analyze cause and effect, and ways to make improvement, if needed. For the preferred suppliers, the analysis will give hints to why the suppliers performed well. These hints will be used to help non-performing suppliers to

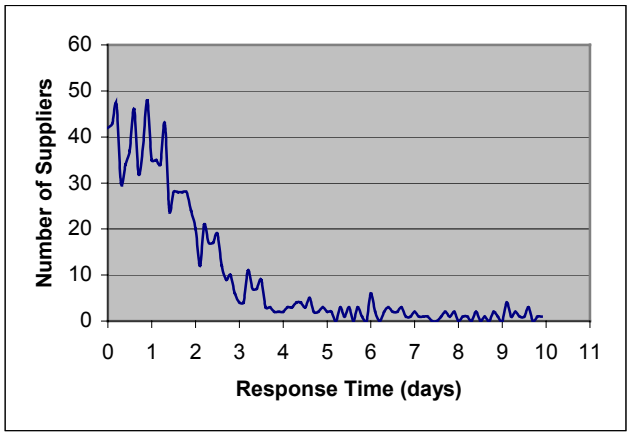

Figure 2. The histogram of the response time of 1000 suppliers.

Supplier Names	Performance	Average	Variance	Contact
ABC Ace	Preferred	0.6	0.3	(800) 123-1000
Houston Metal	Preferred	0.7	0.3	(888) 333-2000
Clear Lake Coins	Preferred	0.8	0.3	(800) 281-2222
Dallas Pencil	Preferred	0.9	0.3	(800) 723-5050
Office Supply	Preferred	1.0	0.5	(800) 312-2828
Sue's Doughnuts	Acceptable	1.5	0.5	(800) 777-5555
Mary Catering	Acceptable	1.5	0.5	(800) 425-2100
Pacific Cable	Acceptable	1.6	0.5	(800) 322-5789
Al's Boot Company	Acceptable	1.7	0.5	(800) 328-4377
Viking Shoes	Acceptable	1.8	0.5	(800) 713-1000
Blue Wing Liqueur	Acceptable	1.9	0.5	(888) 450-1111
Arthur Allen, Inc.	Concerned	2.0	1.3	(888) 444-1200
Red Apple Supply	Concerned	2.0	1.4	(800) 988-1888
Atlantic Shoes	Concerned	2.0	1.5	(800) 382-5600
Bombay's Shoes	Concerned	2.0	1.6	(800) 832-1222
Waterfall, Inc.	Not Acceptable	5.3	0.3	(800) 214-8200
Canadian Coins	Not Acceptable	5.4	1.1	(604) 569-2455
Multi Shoes	Not Acceptable	5.5	2.3	(800) 432-1888
MicroManagement	Not Acceptable	5.6	3.0	(800) 322-8000

Figure 3. Report listing suppliers according to their performance level.

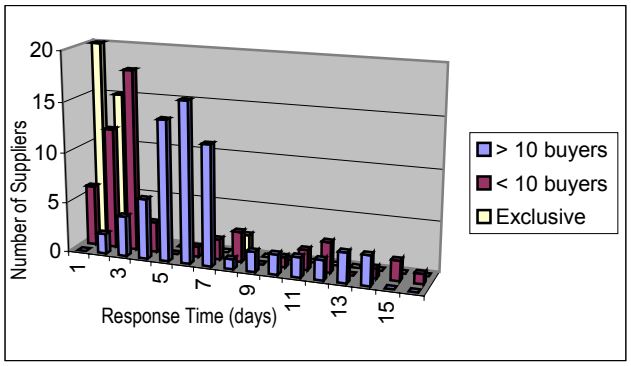

Figure 4. Multi-dimensional histogram of over performance level.

improve their performance. For the non-performing suppliers, the analysis will give hints to how to improve its performance in some periodic review, what corrective measures to suggest and apply, and clarifications on the checklist that determines the approval criteria for suppliers. Figures 5, 6, 7 displays typical data visualization for each individual supplier for this analysis purpose.

C. *Future Task: Automatic Analysis & Correlation*

At this timeframe, the results of the above work were manually analyzed for managerial decision concerning

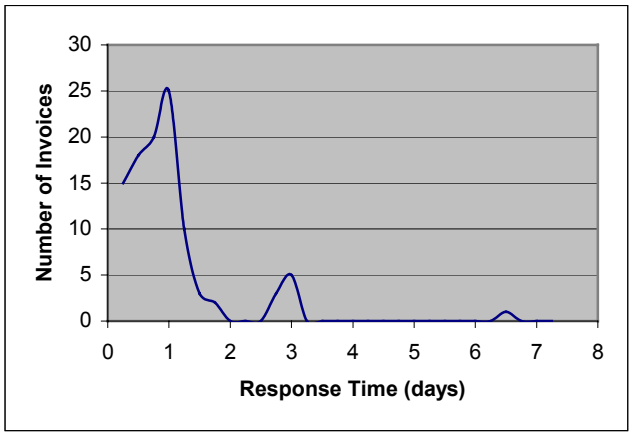

Figure 5. The histogram of the response time of one supplier.

Invoice Number	Invoice Date	Performance	Resp. Time	Seller Name
2003030200001	03/02/2003	Preferred	1	John Doe
2003031800033	03/18/2003	Preferred	1	Jack Brown
2003032000182	03/20/2003	Preferred	1	Jack Brown
2003032100313	03/21/2003	Preferred	1	John Doe
2003032700927	03/27/2003	Preferred	1	John Doe
2003030500011	03/05/2003	Acceptable	2	Jack Brown
2003030600012	03/06/2003	Acceptable	2	Jack Brown
2003030600013	03/08/2003	Acceptable	2	Jack Brown
2003030600014	03/10/2003	Acceptable	2	John Doe
2003030600015	03/10/2003	Acceptable	2	John Doe
2003030600016	03/11/2003	Acceptable	2	John Doe
2003030600017	03/12/2003	Concerned	3	Jane Doe
2003030600018	03/12/2003	Concerned	3	Jane Doe
2003030600019	03/12/2003	Concerned	3	Jane Doe
2003030600020	03/12/2003	Concerned	3	Jane Doe
2003030600021	03/15/2003	Not Acceptable	5	Jane Doe
2003030600022	03/15/2003	Not Acceptable	6	Jane Doe
2003030600023	03/15/2003	Not Acceptable	6	Jane Doe
2003030600024	03/15/2003	Not Acceptable	9	Jane Doe

Figure 6. Report listing individual performance level.

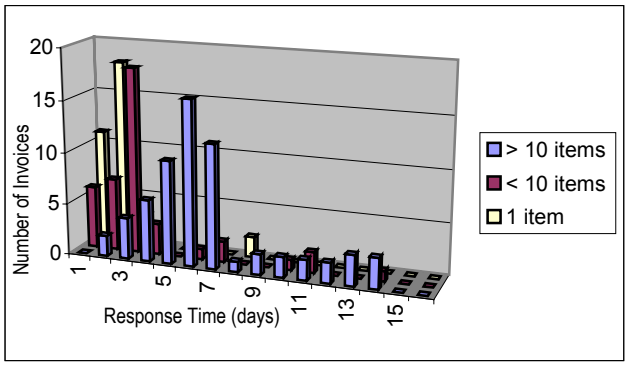

Figure 7. Multi-dimensional plot of individual performance level.

individual suppliers. This task is tedious and sometimes can be subjective. The number of dependent variables involved, when being large, is another obstacle for manual analysis because of the difficulty in data visualization in multi-dimensional space.

V. Conclusion

It has been shown that raw EDI data can be mined for practical application of measuring the performance of a supply chain. The EDI data are parsed into individual data elements, the performance is calculated for each supplier,

and the final results are presented as the whole for the supply chain as well as individually for each supplier for performance review and for planning control action. Numerical examples are presented to demonstrate the workability of the concept and to pose the definition of future work on analysis of the result of this data mining process.

REFERENCES

[1] H. A. Abbass, R. A. Sarker, and C. S. Newton (eds.). *Data mining: a heuristic approach*. Hershey, PA: Idea Group (2002).

[2] R. Groth. *Data mining: a hands-on approach for business professionals*. Upper Saddle River, NJ: Prentice Hall PTR (1998).

[3] J. Han and M. Kamber. *Data mining: concepts and techniques*. San Francisco, CA: Morgan Kaufmann Publishers (2001).

[4] N.F.F. Ebecken (ed.). *Data mining*. Boston, MA: WIT Press Computational Mechanics Publications (1998).

[5] B. Thuraisingham. *Data mining: technologies, techniques, tools, and trends*. Boca Raton, FL: CRC Press (1999).

[6] A.D. Gordon. *Classification: methods for the exploratory analysis of multivariate data*. New York, NY: Chapman and Hall (1981).

[7] J. W. Tukey. *Exploratory data analysis*. Reading, MA: Addison-Wesley Pub. Co. (1977).

[8] K. Jajuga, A. Sokolowski, H. H. Bock (eds.). *Classification, clustering, and data analysis: recent advances and applications*. New York, NY: Springer (2002).

[9] B. W. Silverman. *Density estimation for statistics and data analysis*. New York, NY: Chapman and Hall (1986).

[10] M. James. *Classification algorithms*. London, UK: Collins (1985).

[11] R. F. Gunst, R. L. Mason. *Regression analysis and its application: a data-oriented approach*. New York, NY: M. Dekker (1980).

[12] Michael J. Cunningham. *B2B: How to build a profitable e-commerce*. Cambridge, MA: Perseus (2001).

[13] Steve Lohr. "Business to business in the Internet." *The New York Times*. (April 28, 1997).

[14] Warren D. Raisch. *The eMarketplace: strategies for success in B2B eCommerce*. New York, NY: McGraw-Hill (2001).

[15] D. E. O'Leary. *Enterprise resource planning systems: systems, life cycles, electronic commerce, and risk*. Cambridge, UK: Cambridge University Press (2000).

[16] R. A. Miranda (ed.). *ERP and financial management systems: the backbone of digital government*. Chicago, IL: Government Finance Officers Association (2001).

[17] Data Interchange Standards Association. *Electronic data interchange X12 standards: draft version 4, release 3*. Alexandria, VA: Data Interchange Standards Association (1999).

[18] J. J. Martin. *Data types and data structures*. Englewood Cliffs, NJ: Prentice-Hall (1986).

[19] D. A. Tugal, and O. Tugal. *Data transmission*. New York, NY: McGraw-Hill (1989).

[20] M. Sloman (ed.). *Network and distributed systems management*. Reading, MA: Addison-Wesley (1994).

[21] T. Pham, and R. J. P. deFigueiredo. "Maximum Likelihood Estimation of a Class of Non-Gaussian Densities with Application to l_p Deconvolution" *IEEE Journal in Acoustics, Spectral, and Signal Processing* (Jan 1989).