Volume 4, Number 1 May 2006

The International Journal of Applied Management and Technology

A refereed journal published by the School of Management, Walden University

The International Journal of Applied Management and Technology (iJAMT) is published biannually in May and November. Publication is via digital media and available for viewing or download from the journal's web site at http://www.ijamt.org.

THE INTERNATIONAL JOURNAL OF APPLIED MANAGEMENT & TECHNOLOGY
IJAMT

ISSN 1554-4740
www.ijamt.org | info@ijamt.org

Submissions are welcomed from scholars, scholar-practitioners, and advanced graduate students. Please visit the journal's web site at www.ijamt.org for detailed information.

The International Journal of Applied Management and Technology is a peer reviewed publication.

Advisory Board

Editorial Board

Contents

A MODEL FOR DETERMINING THE QUALITY OF THE MANUFACTURING SYSTEM A USING NEURO-FUZZY APPROACH

A. Noorul Haq, National Institute of Technology, India
G. Kannan, K. L. College of Engineering, India
N. Gunasekaran, Kumaraguru College of Technology, India

ABSTRACT

Though India is the largest producer of sugar it does not rank among the best in terms of sugar quality. Most of the sugar industries use conventional control techniques, which found fall short of providing effective means of controlling the processes when the process is complex. With the increasing demand for better quality and productivity, there is a need to automate the manufacturing processes. Many processes are however complex, non-linear, stochastic and even ill defined in some cases. In fact operation of many processes still relies on the operator's skills due to lack of robust control methods. However, it is well known that humans, once trained or experienced, can successfully control various complex systems without having to depend on mathematical models. This fact leads to the need to develop intelligent monitoring and control schemes.

In this paper, effort has been made for intelligent optimization and control of complex process to achieve better quality of sugar using intelligent control schemes based on Fuzzy Logic and Neural network .An attempt has been made to study the entire process of sugar, various factors that affect the quality of sugar, existing control methods and its effect on final product. The study aids to develop a new model for underlying non-linear dynamic process using fuzzy logic and Neural Network. The paper is a systematic attempt to design knowledge based Neuro-fuzzy model for a complex system to make the plant to operate safely and efficiently. The above intelligent control schemes overcome the limitations of conventional control methods for complex problems. The effectiveness of the model is illustrated using a sugar and beverages manufacturing company, located at southern part of India and validated the results using Fuzzy logic and Neural network.

Keywords

Fuzzy Logic, Neural Network, Sugar Company

Introduction

Today sugar is the most widely used sweetener. It is most often used to sweeten foods and beverages. Sucrose is generally known as cane sugar, which is used as a sweetening agent for foods as in manufacture of candies, cakes, puddings preserves, soft and alcoholic beverages and many other foods. Sucrose a basic foodstuff, supply approximately 13 % of all energy that is derived from foods. Sucrose is a disaccharide that it is made up of two simple sugars namely glucose and fructose.

The extraction of sucrose from cane sugar is a complex process, which undergoes several treatments before manufacturing final product. Though India is the largest producer of sugar, it does not ranks among the best in terms of sugar quality, since most of the sugar mills in India follows traditional method of manufacturing sugar which lacks in monitoring and in process control techniques. In order to produce sugar of superior quality that would be saleable in international market a cost effective route and modern process control techniques have to be implemented to produce quality sugar. The conventional method of juice extraction suffers from drawback of pH control of juice, control of limewater, sulphurdioxide, impurities in juice, dextrons (bacterial content), control of temperature, pressure, and flow of juice etc which affects the quality of juice.

Advanced process control technique like Neuro-fuzzy controller can be used to control the sugarcane process effectively. Improper control of pH of limewater, and sulphurdioxide not only affects the quality of sugar but also leads to abnormal condition of sugar plant. Therefore information on abnormal events should also be considered important for risk assessment in process industries. Thus the operator is forced to focus through large amounts of interrelated data to identify the causes, which leads to poor quality of sugar. This requires integration and analysis of large amount numerical and abstract data. Therefore a knowledge based artificial intelligence system is more appropriate for such situations. Fuzzy and Neural Networks separately and in combination, plays an increasingly important role in understanding human cognition and to simulate human decision making under uncertain and imprecise environments. They learn from experience with numerical and linguistic sample data. Learning from samples involves construction of a model of the system from the knowledge of collection of input-output pairs.

The need and scope of the paper deals with lack of experience in modeling on the dynamics of the process industry. Lot of training is required for the proper maintenance of the system and continuous improvement of quality which is incorporated with the process industry. The main aim of the paper is to build a model for the manufacturing system, to have an experience of modeling to integrate the Neuro- Fuzzy computing in the process manufacturing system and to analyze how the quality is affected or changed with the simulated model.

Literature survey

In this paper the more importance is given to improve the quality of sugar using advanced process control techniques and hence the initial phase of the study begins on conducting literature survey on process details of cane juice, purity of juice, sucrose content in juice, brix level, factors that affect the quality of juice and final product, impact of agricultural factors on quality, effect of low and high pH of juice, process control techniques etc.

The detailed analysis of cane juice, to test and control the pH value, purity, sucrose content, brix percentage, dextrons, sulphur content, color of the juice, moisture content, grades of the sugar are found in Verma and the formula for calculation of purity of juice pol (sucrose) percentage, brix percentage, moisture content, ash content and dextron percentage which helps to determine the quality of sugar (Verma, 1998).

According to Legendre et al (1998) the quality of sugar not only depends on process control technique but also varies according to the purity of cane juice .The purity of juice, sucrose content in juice, presence of impurities in juice, dextrin (bacterial content), brix of juice depends on agricultural factors like presence of nutrients in soil, pH value of soil, deficiency of nutrients, climatic conditions, post harvesting of cane, trash content etc. (Legendre, et al, 1998).

The method of separating sugar and molasses, removing impurities, developing sugar crystals, grain size, color and grading of the sugar is explained in the book by Chung et al (1996). The poor control of pH of cane juice and addition of excess or deficient supply of limewater and sulphurdioxide gas affects the quality of juice and which reduces the nature of quality of sugar. Low pH or high pH due to excess addition of sulphurdioxide or due to excess addition of lime water leads to transient conditions of the plant by affecting the entire system . The literature was also collected from the book "Hand book of Industrial water conditioning" by Betz (1980). The details about pH, effects of low pH and high pH is collected from the book Jain and Jain (1980).

The definition of fuzzy logic, its principles, application of fuzzy logic in control and signal processing was given by Zadeh (1965). According to Zadeh fuzzy logic is all about the relative importance of precision. As complexity arises precise statement loses meaning and meaningful statements lose precision. Based on Zadeh fuzzy logic is a fascinating area of research because it does a good job of trading between significance and precision. In some critical decision making situations high degree of fuzziness and uncertainties are involved in the data set. Fuzzy set theory provides a framework for handling the uncertainties of this type.

According to Bart and Kosko (1997) the neural networks and fuzzy systems estimate functions from sample data and also used adaptive rule tuning for output and input variable problems. According to Munakata and Jain (1994) fuzziness means multivaluedness or multivalence. Multivalued fuzziness corresponds to degrees of uncertainity, partial occurrence of events or relations. Fuzzy logic can be used in control systems to control the process effectively.

According to Jang and Sun (1995) the process can be further effectively controlled by developing Neuro – fuzzy controller, where the neural network system train the network according to the dynamic changes in the behaviour of the system. Once the network is trained it initiates the fuzzy controller to take appropriate control actions. Several authors have explained the above control methods for solving various problems in different fields.

Shin and Vishnupad (1996) discuss that operation of many complex processes still depends on operators skill due to lack of robust control methods. Humans once trained or experienced can successfully control such complex systems .The author stress that fact and emphasis that there is a need to develop an intelligent monitoring and control schemes for the non-linear process.

The basic idea behind the proposed scheme is first to capture the non-linear relationship between the operating parameters (control variables) and the resultant process conditions from empirical data which is collected during normal operation using the excellent modeling capability of neural networks and then to use the model based fuzzy control scheme to determine the control output.

According to Shin and VishNupad (1996) a model based Neuro-fuzzy optimization has been demonstrated as an effective means of achieving non-linear optimization of complex problems. Based on current values of input fuzzy rules are generated from the neural network and membership functions are assigned to the variables. It has been recently reported that advanced process control technique can improve the sugar yield, reduce energy consumption, increase capacity, improve sugar quality and consistency, improve process safety and reduce environmental emissions.

Bellman and Zadeh (1970) presented some applications of fuzzy theories to the various decision-making processes in a fuzzy environment. Zimmerman (1976, 1978) presented a fuzzy optimization technique to linear programming (LP) problem with single and multiple objectives. Since then the fuzzy set theory has been applied to formulate and solve the problems in various areas such as artificial intelligence, image processing, robotics, pattern recognition, etc. Yang, Ignizio, and Kim (1991) formulated the FGP with nonlinear membership functions.

Problem description and Analysis

Though India is the largest producer of sugar and sugarcane it does not ranks among the best in terms of sugar quality. The main reason that India couldn't able to produce sugars to meet international standards are due to failure in implementing advanced process control techniques. The major problem in sugarcane process industry while processing is improper control of pH of cane juice, milk of lime, sulphur-di-oxide, presence of impurities in cane juice, grain size of sugar, suspended matters and color of the syrup etc affects the quality of sugar.There are several other non-technical factors like soil, nutrients, cane varieties climate, Trash, sucrose and non-sucrose content, nature of juice etc which also affects the quality of sugar.

The above problem can be solved by advanced control techniques like Neuro-fuzzy controller along with Simulink model. A Neuro-fuzzy model have been built up using matlab to show the effect of pH, lime water, sulphurdioxide, impurities and color of juice on the quality of the final product. The quality of sugar manufactured should meet the international standard ICUMSA (International Commission for Uniform Methods of Sugar Analysis), which is a worldwide body that brings together the activities of the National Committees for Sugar Analysis in more than thirty member countries.

Factors affecting the quality sugar

a. Nature of Juice
b. Presence of non-sucrose in cane.
c. pH of juice.
d. Excess or deficient supply of lime water and sulphurdioxide.
e. Brix of lime.
f. Brix of juice.
g. Adequate or deficiency of nutrients in soil.
h. Color of the juice.
i. Temperature and flow control of juice.
j. Delay in harvesting and crushing.
k. Loss of sucrose in cane.
l. pH of soil.
m. Grade and grain size of sugar.
n. Conventional process control methods.
o. Bacterial contents in juice.

Parameters used to measure the quality of sugar

There are several parameters to be measured based on which the quality of the sugar can be tested, graded and confirmed whether it has to be accepted or not. Some of the parameters to be measured are

I. Grade of the sugar as per ISS
   a. L 31 M 31 S 31
   b. L 30 M 30 S 30
   c. L 29 L 29 S 29

II. Grade of the sugar as per ICUMSA
   a. 100 – Standard
   b. Between 80 to 90 – Good Quality
   c. Between 40 to 70 – Superior Quality
   d. Between 100 to 125 – Accepted Level
   e. Greater than 125 – Inferior Quality

III. Grain size
   a. L – Extracourse (> 1mm)
   b. M – Standard (0.25 TO 0.75 mm)
   c. S – Small (Or) Fine Grain Size (< 0.25mm)

IV. Color of the sugar
   a. Pure White – Superior Quality
   b. Dull White – Good
   c. Dark Brown – Inferior

V. Presence of Sulphur
   a. 0 to 16 ppm – Accepted
   b. 0 – Good
   c. > 16 ppm – Inferior

VI. Presence of Lead – Nil

VII. Presence of Ash
   a. 0 to 0.05% – Accepted
   b. > 0.05% – Inferior
   c. 0 – Good

VIII. Presence of Moisture
   a. 0 to 0.04% – Accepted
   b. 0 – Good
   c. >0.05% – Inferior

IX. Presence of Dextron
   a. 0 to 100 ppm – Accepted
   b. 0 ppm – Very Good
   c. > 100 ppm – Inferior

X. Presence of Sediments
   a. 0 to 100 ppm – Accepted
   b. 0 ppm – Very Good
   c. > 100 ppm – Inferior

Solution methodology

The modeling tool such as Fuzzy logic and Neural network is adopted to solve the problem.

Introduction to fuzzy logic system

Fuzzy logic was invented by LOFTIZADEH. Fuzzy logic is a generalization of classical logic in which there is a smooth transition from true to false. Fuzzy uses numeric computation between 0 to 1 to facilitate approximate reasoning.The basics of fuzzy logic are derived from the fuzzy set theory. The fuzzy set is defined in such a way to dichotomize the individuals in some given universe of discourse in two groups as members (those that certainly belongs to the set) and non-members (those that certainly not belong to the set)

A fuzzy set A is X is characterized by a membership function (MF) $\mu_A(x)$, which associates a real number with each element in X in the interval [0,1] (with the value of $\mu_A(x)$ representing the grade of membership of X in A). Thus the fuzzy set A on a universe of discourse X is defined as

$$A = \{(x, \mu_A(x))/ x \varepsilon X\} \qquad \text{-----------------(1)}$$

The MF is a curve that maps each element to its membership value between 0 and 1.There are various types of membership functions. They are of triangular type, trapezoidal, Gaussian, generalized bell shaped, S-Shape, Z-Shape and Sigmoidal shaped functions.

A fuzzy set can be defined mathematically by assigning each possible individual in the universe of discourse, a value representing its grade of membership in fuzzy set. The individuals may belong to the fuzzy sets to a greater or lesser degree as indicated by a larger or smaller membership grade. These membership grades are very often represented by real number values ranging in the closed interval between 0 and 1.

Fuzzy logic belongs to the class of knowledge-based system. Fuzzy logic can mimic experts. Fuzzy systems are knowledge based software environments constructed from a collection of linguistic If-Then rules. Fuzzy systems realize non-linear mappings and hence they are recognize as alternative control schemes for dynamic, non-linear and complex process.

The past few years have witnessed a rapid growth in the number and variety of application of fuzzy logic. The application ranges from consumer products, refrigerators, washing machines, microwave ovens, medical instrumentations, decision support systems and in industries. To understand the reasons for the growing use of fuzzy logic it is necessary first to clarify the meaning of fuzzy.

The basic concept underlying fuzzy logic is that of a linguistic variable (i.e.) a variable whose values are words rather than numbers. Another basic concept in fuzzy logic, which plays a central role in most of its applications, is that of a fuzzy IF-THEN rule.

In rule based fuzzy logic, machinery is provided by calculus of fuzzy rules. The calculus of fuzzy rules serves as a basis for what might be called the fuzzy dependency and command language (FDCL).

In most of the applications of fuzzy logic, a fuzzy logic solution is in reality a translation of a human solution in to fuzzy dependency and command language. Fuzzy logic is viewed as soft computing. Unlike the traditional hard computing, soft computing is aimed at an accommodation with the pervasive impression of the real world. The guiding principle of soft computing is to exploit the tolerance for impression, uncertainty and partial truth to achieve a) traceability b) robustness and c) low solution cost.

Among various combinations of methodologies in soft computing, the one that has highest visibility at this juncture is that of fuzzy logic and neuro-computing leading to so called neuro fuzzy systems.

Fuzzy system

A fuzzy system is a non-linear mapping of an input data vector in to a scalar output using a fuzzy logic. This mapping is performed using the fuzzification, fuzzy inference and defuzzification components and is shown in Fig 1

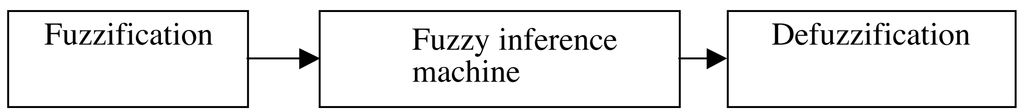


Fig1: Structure of fuzzy system

Fuzzification

The fuzzification component maps an observed input space into suitable linguistic values which can be viewed as labels of fuzzy set.

Fuzzy inference machine

The fuzzy inference machine consists of two conceptual parts.

i. A rule base which contains a selection of fuzzy rules. "If-Then" rules are used to relate fuzzy variables.
ii. A database defining the membership functions used in the fuzzy rules, normalizing the input and output universe of discourse and performing the fuzzy partition of input and output spaces. A reasoning mechanism which performs the inference procedure upon the rules and given conditions to derive a reasonable output. A simple fuzzy rule has the form if x is A, y is B, where A and B are the input and output linguistic variables defined by fuzzy sets on the universe of discourse X and Y respectively. In the if part A is called the antecedent and the other part B is called the consequence. The rule can be represented as a fuzzy relation as $R = A \rightarrow B$, where R is the fuzzy relationship defined on the cartesian product universe X*Y. A and B are the antecedent part and the consequent part respectively. These relations may be obtained from sources such as human expert knowledge and experience, commonsense and engineers intuitive knowledge about the system under development, general principles and laws governing the system, by using pattern recognition classification, clustering and learning techniques. The process to obtain the inference result from the observed data is predominantly used max-min compositional technique.

Defuzzification

It converts the aggregated fuzzy set to a non-fuzzy output value by taking the average of the area bounded by the membership function curve.

Structure of Fuzzy System

The structure of the fuzzy system is shown in Fig 2.

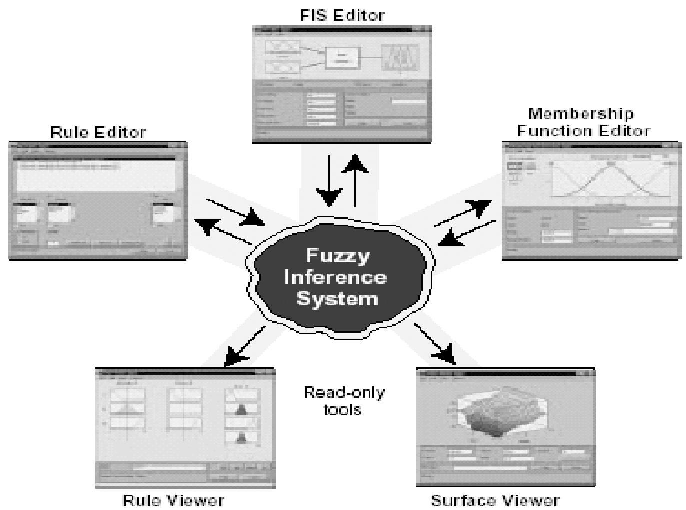


Fig2: Structure of Fuzzy System

Overview of Artificial Neural Network

In recent years artificial neural networks (ANNs) attracted more attention from both academic researchers and industrial practitioners. Artificial neural networks, powerful pattern recognition and flexible non-linear modeling capabilities are the main reasons for its popularity. Neural networks offer great promise for process control. Advanced manufacturing systems are becoming too complex and dynamic for traditional ruled based decision support systems. Neural network technology application is gathering momentum in all aspects of manufacturing processes. Since decision making process in an advanced manufacturing systems environment is becoming increasingly difficult and overwhelming to humans, Artificial intelligence (AI) is becoming widely adopted to assist human efforts. There is a wide use of several aspects of AI in all aspects of manufacturing processes. The most common type of AI is expert systems (ES).But expert system are less effective in the ever changing , complex and open system environment of today's manufacturing system.

The type of AI capable of responding to the rapid changes and ill structured nature of the automated manufacturing environment has to be able to learn, adapt to changes, recognize trends and mimic human thoughts. It must be able to recognize patterns and relationships among prevailing factors which are not apparent to human experts. Neural net technology fits the above to a greater extent. Neural net's ability to recognize patterns and trends in an historic database is one of the reasons that the technology is effectively used in manufacturing processes.

Design of Neural network

The design issues of neural networks are so complex. Designing of network consist of

a. Arranging neurons in various layers.
b. Deciding the type of connection among neurons for different layers as well as among neurons with in a layer.

c. Deciding a way neuron receives input and produces output.
d. Determining the strength of connections with in the network by allowing the network, the approximate values of connection weights by using a training data set.

Layers

Biologically neural networks are constructed in a three dimensional way from microscopic components. These neurons are found to be of unrestricted interconnections. Artificial neural networks are the simple clustering of the primitive artificial neurons. In artificial neural networks neurons are grouped in to three layers.

- ❖ The input layer that receives input data from the external environment.
- ❖ The output layer consists of neurons that communicate the output of the system to the user or external environment.
- ❖ Between the input and output layer there are number of hidden layers.

A neuron receives input signals and in turn, passes a single signal to other neurons through weighted connection feeding in to summation box. The output signal passes through several pathways to provide input signals to other neurons. The connection between any two neurons is called synapse, through which neurons send and receive messages from each other. Some of the connections are stronger than other based on the adaptive coefficients often referred to as weights.

Learning

The brain basically learns from experience. Neural networks are called as machine learning algorithms, because changing of its connection weights (training) causes the network to learn the solution to the problem. The strength of the connection between the neurons is stored as a weight, which is nothing but a particular value for specific connection. The system learns new knowledge by adjusting this connection weights. The learning ability of a neutral network is determined by its architecture and by the algorithm method chosen for training.

The two forms of learning in neural networks are supervised and unsupervised. In supervised training a trainer is needed to specify the desired output for a given input. The training process involves providing the network with sets of inputs and the corresponding outputs. A network is said to have learned if it can produce the desired outputs when a sequence of inputs are provided.

Unsupervised learning involves the trainer informing the network of the input data and of the number of output parameters and the system that figures out the solution.

Overview of the company

The research was conducted at sugar and beverages manufacturing company located at southern part of India. After careful consideration of many industries for the purpose of study, this sugar mill is selected for conducting this study. At present this concern is premier in the manufacture of high quality white crystal sugar and industrial alcohol. Sakthi sugars limited is one of the largest producers of sugar in India having five units, three units in Tamilnadu and two units in Orissa states accounting for about 15,000 tonnes of cane crushed per day. The quality of its sugar has been widely acclaimed as evident from the premium it has been getting consistently. The sugar recovery has witnessed a remarkable improvement due to the efforts taken by the company in cane research and development.

Constructing a new neuro-fuzzy model

A new application of neuro-fuzzy system to the processing of sugar is developed using Matlab software package. The first step in constructing a model is to observe the real sugarcane processing system and the interaction among its various components and system behaviour. The second step in model building is construction of a conceptual model by collection of various data's and the structure of the system. The third step is the translation of the operation model in to computer recognizable format. While building a model there is a need for continuous verification and validation by comparing the real system to the conceptual model and repeated modification is done to improve its accuracy.

Three fuzzy models of type mamdani are developed by mapping an input data vector into a scalar output using fuzzy logic. In first model pH of crushed juice is taken as input and lime water to be added to juice is taken as output and the mapping between these data which are non-linear mapping are performed using fuzzification, fuzzy inference and defuzzification components found in fuzzy logic tool box in matlab. Similarly in the second model pH of juice after treatment of lime water is taken as input and sulphurdioxide to be added to juice is taken as output and the data are mapped in a similar manner and in the third fuzzy model the parameters which is used to test the quality of sugar is taken as input and the quality of sugar is taken as output and the data are mapped in a non-linear manner.

Similarly the fuzzy logic tool box is highly impressive in all respects. It makes fuzzy logic an effective tool for conception and design of intelligent system. The fuzzy logic toolbox is a collection of functions built on the matlab with numeric computing environment. It provides tool to create and edit fuzzy inference systems. The fuzzy system can be integrated with simulink and the model can be simulated. The fuzzy models can be created using graphical tools or by using command line functions.

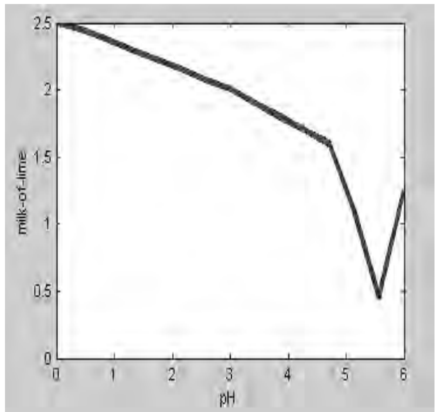
Model : 1a Treatment of ph of juice with lime water (Fuzzy)

Input : pH of mixed juice.
Output : Milk of lime

TABLE 1: Input and output Membership functions label for model 1a

S.No	Membership functions	Input		Output	
		Membership functions label	Range	Membership functions label	Range
1	mf1	VHA	0 – 3	VLL	0-0.6
2	mf2	HA	3 – 4.7	LL	0.6-1.0
3	mf3	MA	4.7 – 5.2	ML	1.0-1.6
4	mf4	LA	5.2 – 5.4	HL	1.6-2.0
5	mf5	VLA	5.4 – 6.0	VHL	2.0-2.5

Fig3: Fuzzy Output for the Model: 1a



Model : 1b Treatment of pH of juice with lime water (Neural)

Input : pH values of cane juice before lime treatment
Target : Different values of Lime water collected during process study

Training Details

Network type : Linear (train) neural network
Input range : 0.0549 to 5.98
No of Neurons : 1
Input delay vector : 0
Learning rate : 0.001
Initial weight : 0
Adopted weight : 0.060652
Initial bias : 0
Adopted bias : 0.02458

Training parameters

No of epochs : 7912
Goal : 0
Max-fail : 5
Show : 100
Time : Inf

TABLE 2: Input, target values, output and error value of network model 1b .

Input (pH of juice)	Target values	Output	Error =(Target – Output)
0.0549	2.5	2.5353	- 0.035272
0.53	2.43	2.4457	- 0.015708
1.48	2.3	2.2666	0.033382
1.85	2.2	2.1969	0.00031333
2.36	2.13	2.1007	0.0029603
2.95	2.01	1.997	0.013
3.05	1.98	1.9706	0.0093526
3.25	1.95	1.9326	0.017056
3.57	1.88	1.8726	0.007381
3.82	1.83	1.8255	0.00451
4.26	1.73	1.7425	- 0.012543
4.52	1.65	1.6935	- 0.043529

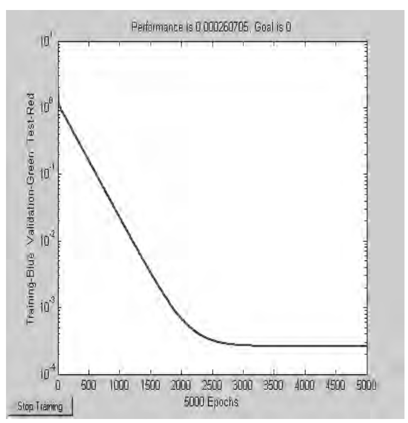


Fig4: Neural Output for the Model :1b

MODEL : 2a Treatment of ph of juice with Sulphur dioxide

Input : pH of mixed juice. (Base)

Output : SO_2 gas

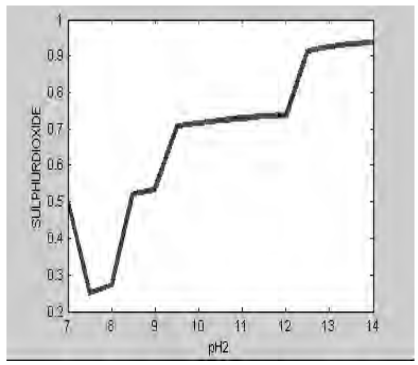


TABLE 3: Input and output Membership functions label for model: 2a

S.NO	Membership functions	INPUT		OUTPUT	
		Membership functions label	Range	Membership functions label	Range
1	mf1	Low-alk	7-8	VLS	0-0.4
2	mf2	Medium-alk	8-9.2	LS	0.4-0.6
3	mf3	High-alk	9.2-12	HS	0.6-0.8
4	mf4	VHA	12-14	VHS	0.8-1.0

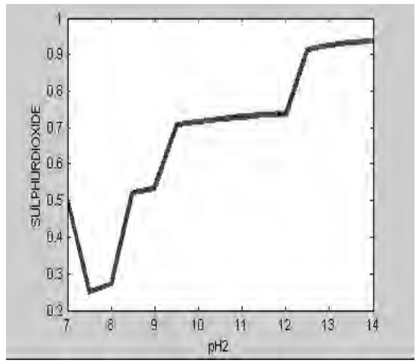


Fig5: Fuzzy Output for the Model: 2a

MODEL: 2b Treatment of pH of juice with Sulphur dioxide

Input	: pH values of cane juice after lime treatment
Target	: Different values of sulphur dioxide collected during process study

Training Details

Network type	: Linear (train) neural network
Input range	: 0.7.25 to 11.8
No of Neurons	: 1
Input delay vector	: 0
Learning rate	: 0.001
Initial weight	: 0
Adopted weight	: 0.064518
Initial bias	: 0
Adopted bias	: -0.041059

Training parameters

No of epochs	: 10000
Goal	:0
Max-fail	: 5
Show	:125
Time	:Inf

TABLE 4: Input, target values, output and error value of network model

Input (pH of juice)	Target values	Output	Error =(Target – Output)
7.25	0.225	0.30158	- 0.076575
7.36	0.236	0.31545	- 0.079447
7.53	0.25	0.33689	- 0.086886
7.83	0.266	0.37472	- 0.10872
7.92	0.269	0.38607	- 0.11707

8.05	0.505	0.40246	0.10254
8.34	0.516	0.49902	0.01698
8.73	0.529	0.52418	0.00482
8.77	0.53	0.52976	0.00024
8.9	0.533	0.53314	- 0.00014
8.98	0.534	0.53614	- 0.00214
9.11	0.536	0.5440	- 0.008
9.24	0.705	0.59037	0.15247
9.54	0.709	0.61685	0.09215
9.75	0.712	0.66099	0.05101
10.1	0.717	0.69882	0.01818
10.4	0.722	0.72449	- 0.00249
11	0.73	0.73456	- 0.00456
11.4	0.733	0.69445	0.038554
11.8	0.736	0.72025	0.01574

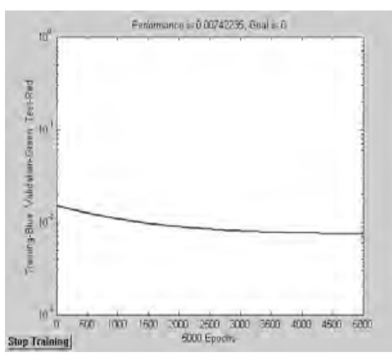


Fig6: Neural Output for the Model: 2b

MODEL: 3 To determine the quality of sugar

Inputs :

a. Grain
b. Color
c. Ash
d. Moisture
e. Dextron

TABLE 5: Input and output Membership functions label for model: 3

Label	Parameter	Membership functions	Membership functions label	Range
Input	Grain	mf1	fine	0 - 0.25
		mf2	standard	0.25 - 1
		mf3	extracourse	1- 1.5
	Color	mf1	purewhite	40 - 60
		mf2	white	60 - 90
		mf3	Stand	90 - 100
		mf4	accept	100 - 125
		mf5	unacceptable	125 - 150
	Ash	mf1	acceptable	0 – 0.05
		mf2	unacceptable	0.05 – 0.15
	Moisture	mf1	acceptable	0 - 0.04
		mf2	unacceptable	0.04 – 0.08
	Dextron	mf1	acceptable	0 - 100
		mf2	unacceptable	100 - 200
Output	Quality	mf1	highly-inferior	0- 30
		mf2	inferior	30-60
		mf3	average	60-75
		mf4	good	75-85
		mf5	superior	85-100

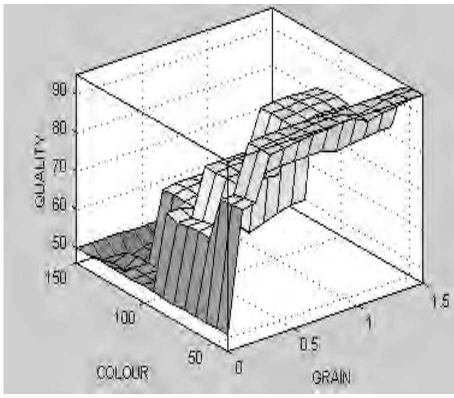


Fig 7(a) : Output with color and grain

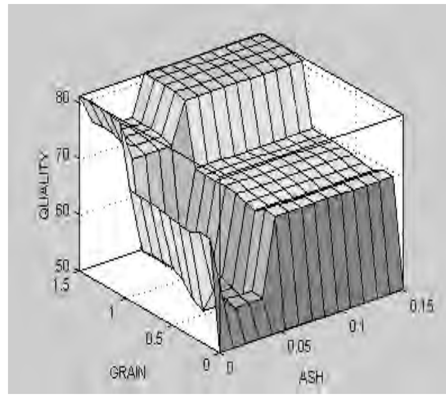


Fig 7(b) : Output with grain and ash

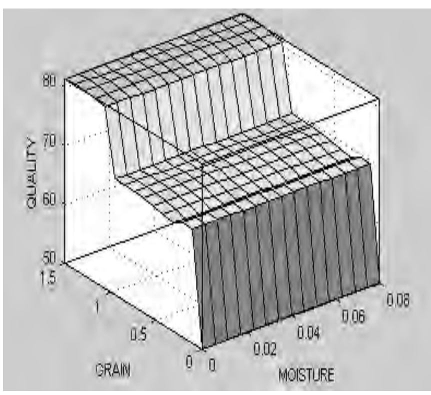


Fig 7(c) : Output with grain and moisture

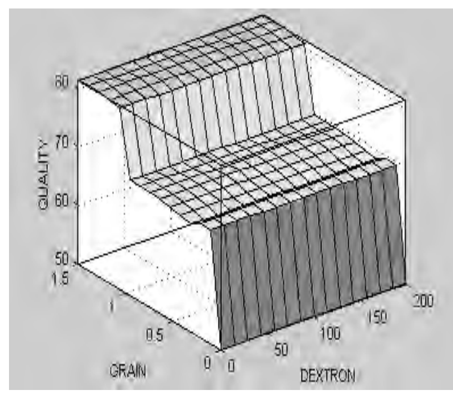


Fig 7(d) : Output with grain and Dextron

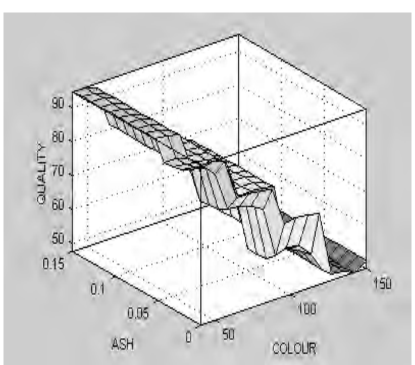


Fig 7(e) : Output with ash and color

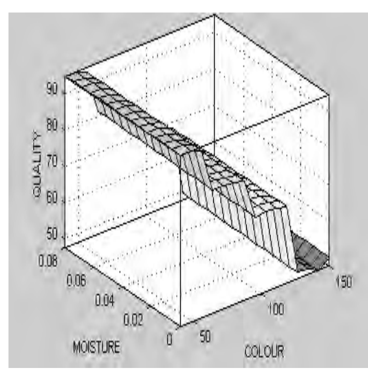


Fig 7(f) : Output with Moisture and color

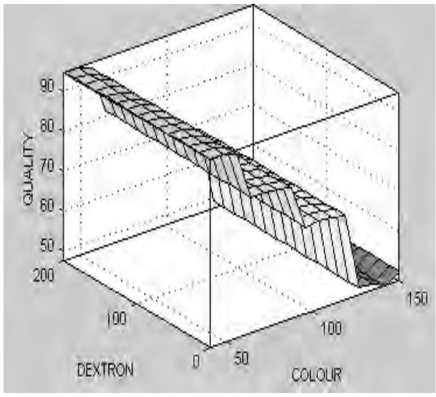


Fig 7(g) : Output with Dextron and color

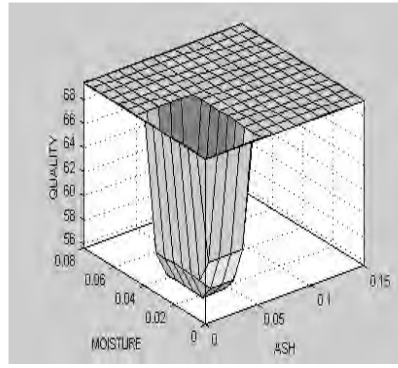


Fig 7(h) : Output with Moisture and ash color

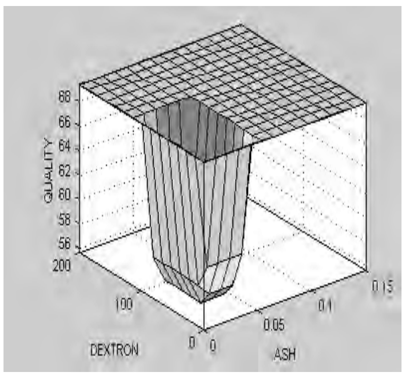


Fig 7(i) : Output with Dextron and ash

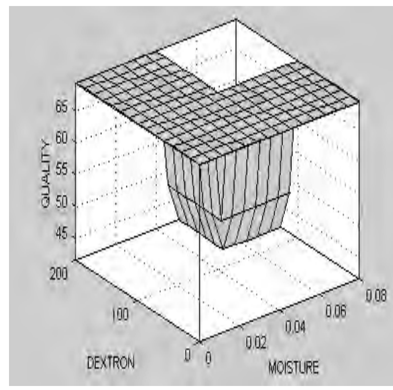


Fig 7(j) : Output with Dextron and Moisture

Fig 7: Fuzzy Output for the Model: 3

Introduction to Neuro-Fuzzy control system

It is generally difficult to accurately model complex systems which are typical characteristics of manufacturing systems. If an accurate analytical model cannot be obtained or the models are incomplete, most quantitative control schemes cannot be applied or even if applicable, the performance of such a control system is not guaranteed. Nonetheless the need to control such complex system frequently arises even when complete models are not available or only partial models are available. Therefore it is desirable to have a generic tool which can be used for modeling complex systems.

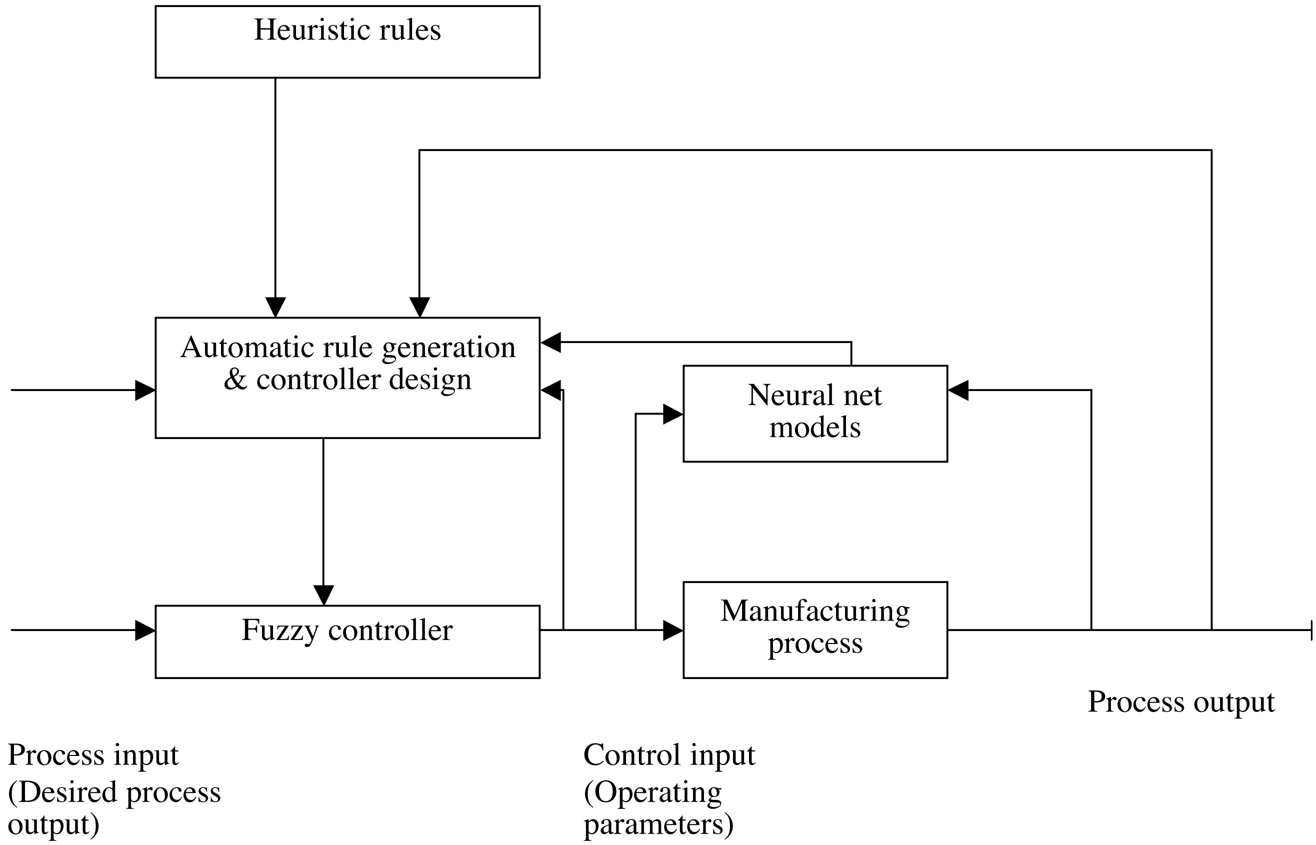
Artificial neural networks have been successfully used to model and approximate various non-linear relationship and systems. Artificial neural networks are trained to learn the mapping between the input and output domains based on observations without requiring knowledge of the underlying system. Neural networks are known for their characteristics of robustness, fault tolerance and generalization. In particular the neural networks have been shown to possess the ability to adapt to dynamic environmental changes through continuous training. Since many control systems are designed on the identified approximately system models, it is also possible to design a controller using a neural network approximating a system. Many different control schemes utilizing neural networks have been proposed to control non-linear systems.

The major advantage of Neuro-fuzzy control over conventional control is that system can be controlled based on input and output observations. Neuro-fuzzy controllers are also cheaper to develop than conventional controllers and can cover a wide range of system operating conditions than conventional controllers. Fuzzy inferencing provides the means of systematically synthesizing various fuzzy rules to produce decision actions so that complex systems can be controlled. Membership functions are used to represent the fuzziness of the observations or actions.

The basic idea behind the proposed scheme is first to capture the non-linear relationship between the operating parameters (control variables) and the resultant process conditions (output variables) from empirical data which can be collected during normal operating using the excellent modeling capability of neural networks, and then to use the model based fuzzy control scheme to determine the control output.

The trained neural networks are used to automatically generate the rules at various operating points, which are subsequently converted in to fuzzy rules and used for fuzzy inferencing. Even if the process is changed dynamically, the neural network can learn the functional relationship between input and output domains through continuous training and the fuzzy controller can derive pertinent control actions. This procedure is similar to the human operator‘s learning process through many observations and repeated trial and errors. As the training process continues, the performance of the system will improve as if an operator can improve skills and knowledge to make decision more intelligently through continuous learning

During processing of cane juice, the operating conditions are continuously monitored and the neural networks are trained for each specific condition. The trained neural model will serve as process models for fuzzy logic inferencing. Based on the difference between desired and monitored process conditions, rules are automatically generated from the neural network for each control variable based on current operating condition and converted in to fuzzy rules with predefined membership functions. Although rules are generated for each pair of input-output variables, the multidimensional relationship of the complex process is not lost and can be fully recovered through fuzzy composition and inferencing. In case of Neuro-fuzzy controllers the control commands are directly generated and automatically tuned by fuzzy inferencing. The Neuro fuzzy control scheme is shown in Fig. 8.



Results and Discussion

In this paper the quality of sugar is determined by considering various factors which have an impact on quality using Neuro fuzzy approach and an advanced Neuro-fuzzy process controller is developed and its control technique is compared with the existing Automatic fuzzy PID pH controller. The Neuro-fuzzy model is developed in matlab using the sample data of pH of juice, milk of lime, sulphurdioxide and other quality testing parameters collected from the process expert in the sugar industry. It has been observed that poor control technique fails to sense the pH of juice accurately and it leads to excess or less addition of lime water and sulphurdioxide which fails to achieve the desired pH value of juice .It was also observed that low pH or high pH of juice not only affects the final quality of the product but also affects the entire operation of sugar plant due to transient conditions of the plant. The collected online process data of critical process variables are trained by neural network by setting target values till the desired values is achieved. Once the Neural network is trained a fuzzy model is created using If-Then rules. By combining Neural network and fuzzy logic a Neuro-fuzzy controller is developed which is an Advanced process controller that overcomes the drawbacks of existing fuzzy PID controller.

The problem is solved by developing a Neuro-fuzzy model for a continuous process using matlab package. This advanced Neuro-fuzzy controller can learn this dynamic behaviour of the process and it trains the network automatically, according to which it controls the addition of limewater and sulphurdioxide gas with juice.

Three fuzzy models are built for continuous process in sugar industry. The result of the first model in Fig3 depicts that for increasing pH value of juice from 0 to 6, the limewater to be added to juice is found to be gradually decreasing. From the above result it was concluded that pH and limewater are inversionally proportional to each other. If pH is 0.3 (High acidity), then high amount of lime, say 2.5g / litre is added to juice to convert pH of acidity to alkalinity. When pH of juice increases then the amount of lime water to be added to juice is found to be in decreasing rate and when pH approaches to 6 milk of lime to be added declines to zero.The declining curve in Fig 3 and Fig 4 shows the relationship between the process variable pH and lime water.

The result of the second model in Fig 5 shows that for increasing pH value of juice from 7 to 14 , the sulphurdioxide to be added to juice is found to be gradually increasing.From the above result it was concluded that pH and sulphurdioxide are directly proportional to each other.If pH is 7.5 (Very Low alkalinity), then very low amount of sulphur dioxide gas is sufficient say 0.2 ppm / litre is added to juice to convert pH of alkalinity to pH neutral. .When pH of juice increases then the amount of sulphurdioxide gas to be added to juice is found to be in increasing rate and when pH approaches to 14 sulphurdioxide gas to be added reaches maximum.The curve in Fig 5 and Fig 6 shows the relationship between the process variable pH and sulphurdioxide gas.

The final fuzzy model is the model built to determine the quality of sugar and to find the relationship between quality and other quality measuring parameters like grain size, color, ash, moisture and dextrin.The result of the final model in Fig 7 shows that these input parameters have serious impact on sugar in terms of sugar quality. Though there are several other quality parameters which are not mentioned in this model due to its complexity nature, grain size and color are considered to be the most important parameters in fixing the quality of sugar.

The Fig 7(a) shows the relationship between three variables in a three dimensional view. It was found that if color is unacceptable and grain size is fine then quality of the sugar is found to be highly Inferior, if the color of the sugar is acceptable and grain size is extra course then the quality of the sugar is found to be average, if the color of the sugar is pure white and grain size is standard then quality of the sugar is Good.The Fig 7(a) clearly shows that the grain size is extra course and the color of the sugar is pure white then the quality of the sugar is Superior.

The Fig 7(b) shows the relationship between grain size, ash and quality of sugar in the three dimensional view. From figure 7 (b) it was found that if the grain size is fine and ash content of sugar is accepted level, then the quality of the sugar is inferior. Also if the grain size is standard and even the ash content is in the acceptable level the quality of the sugar is highly inferior. The Fig 7(c) shows the relationship between grain size, moisture content and quality of sugar in the three dimensional view. From figure 7 (c) it was found that if the grain size is fine and moisture content of sugar is accepted level, then the quality of the sugar is inferior. If the grain size is extra course and the moisture content is in the acceptable level the quality of the sugar is highly good.

The Fig 7(d) shows the relationship between grain size, dextron content and quality of sugar in the three dimensional view. From figure 7 (d) it was found that if the grain size is fine and dextron content of sugar is accepted level, then the quality of the sugar is inferior. If the grain size is extra course and the dextron content is in the acceptable level the quality of the sugar is highly good. The Fig 7(e) shows the relationship between ash, color of the sugar and quality of sugar in the three dimensional view. From figure 7 (e) it was found that if the ash content is un acceptable and color is white, then the quality of the sugar is highly inferior. If the ash content is acceptable and color is pure white, then the quality of the sugar is superior.

The Fig 7(f) shows the relationship between moisture content, color of the sugar and quality of sugar in the three dimensional view. From figure 7 (f) it was found that if the moisture content is unacceptable and color is standard, then the quality of the sugar is highly inferior. If the moisture content is acceptable and color is pure white, then the quality of the sugar is superior. The Fig 7(g) shows the relationship between dextron content, color of the sugar and quality of sugar in the three dimensional view. From figure 7 (g) it was found that if the dextron content is acceptable and color is white, then the quality of the sugar is highly inferior. If the dextron content is acceptable and color is pure white, then the quality of the sugar is superior. The Fig 7(h) shows the relationship between moisture content, ash and quality of sugar in the three dimensional view. From figure 7 (h) it was found that if the moisture content is unacceptable and ash is acceptable, then the quality of the sugar is highly inferior. If the moisture content is acceptable and ash is acceptable, then the quality of the sugar is superior. The Fig 7(i) shows the relationship between dextron content, ash and quality of sugar in the three dimensional view. From figure 7 (i) it was found that if the dextron content is unacceptable and ash is acceptable, then the quality of the sugar is highly inferior. If the dextron content is acceptable and ash is acceptable, then the quality of the sugar is superior. The Fig 7(j) shows the relationship between dextron content, moisture content and quality of sugar in the three dimensional view. From figure 7 (j) it was found that if the dextron content is acceptable and moisture content is acceptable, then the quality of the sugar is highly inferior. If the dextron content is acceptable and moisture content is acceptable, then the quality of the sugar is superior.

Conclusion

This paper is an organized approach to solve the quality problem in sugar industry using advanced process control techniques, where number of process variables involved is high. These process variables are interrelated with each other. A change in one variable may affect the other and may have a serious impact on the final quality of sugar. Therefore the process has to be monitored and controlled by installing advanced automatic process controller. Under this study, various problematic areas are identified in existing method and an advanced controller namely Neuro-fuzzy controller has been developed using available data. The development of neuro-fuzzy controller accurately sense, trains and accordingly it control the process variable, which aids the sugar industry to produce quality sugar. The move towards advanced process control technique namely Neuro–fuzzy controller is definitely a boon to the company since it accurately control the process variables according to the dynamic change in the behaviour of the system.

There is lot of procedures involved in implementing advanced process control technique. It can only run on trial basis till the process engineers are satisfied with its control performance. Once they are satisfied the next step is to interface the automatic pH neuro fuzzy controller to the computer, where the online process can be continuously governed, trained and controlled by the user using the system itself.

References

Bart and Kosko, (1997). Neural networks and Fuzzy systems -A dynamical systems approach to machine intelligence, University of southern California.

Bellman, R. E., and Zadeh, L. A. (1970). Decision making in a fuzzy environment. Management Sciences, Vol. 17, pp.141–164.

Betz (1980). Hand book of Industrial water conditioning, pp. 105-125.

Chung, C. C., Khalid Iqbal., Min, Y. G., Gao, D. W. and Emmanuel Duffaut (1996), An innovative process for white sugar production , New Orleans , Louisiana , USA .

Jain and Jain (1980). Engineering chemistry-Chemistry of engineering materials, pp. 253-387.

Jang, J.S.R., and Sun, C.T., (1995). Neuro –fuzzy modeling and control, proceedings of IEEE.

Legendre B.L., Clarke M.A., Godshall, M.A., and Grisham M.P. (1998). Developments in sugarcane agriculture that affects processing, sugar proceddings, Sanannah, USA, pp. 160-175.

Munakata and Jain, Y. (1994). Fuzzy systems- an overview communication, ACM, Vol. 37, pp. 69-76.

Shin, Y.C., and Vishnupad, P.C. (1996). Neuro-fuzzy control of complex manufacturing processes, International Journal of Production Research,Vol. 34, No. 12, pp. 3291 – 3309.

Verma, N. C. (1988). System on technical control for cane sugar factories in India, The sugar technologies association of India, Kanpur, India.

Yang, T., Ignizio, J. P., and Kim, H. J. (1991). Fuzzy programming with nonlinear membership functions: piecewise linear approximation, Fuzzy Sets and Systems, Vol. 11, pp. 39–53.

Zadeh, L. A. (1965). Fuzzy sets. Information and Control, Vol. 8, pp. 338–353.

Zimmermann, H. J. (1976). Description and optimization of fuzzy systems, International Journal of General Systems, Vol. 2, pp. 209–215.

Zimmermann, H. J. (1978). Fuzzy programming and linear programming with several objective functions, Fuzzy Sets and Systems, Vol. 1, pp. 45–56.

SCHEDULING OF PARTS AND AS/RS IN AN FMS USING GENETIC ALGORITHM

J. Jerald, SASTRA (Deemed University), India
P. Asokan, National Institute of Technology, India
G. Prabhaharan, National Institute of Technology, India
R. Saravanan, Kumaraguru College of Engineering, India

ABSTRACT

In today's manufacturing environments inventories are maintained at lower levels than in the past. These reduced inventories have led to smaller storage systems, which in turn have created the need for quick access to the material being held in storage. Hence automated storage and retrieval systems that has been used in manufacturing, warehousing and distribution applications must provide quick response time to service requests in order to keep the system operating efficiently. This paper attempts to link the operation of the AS/RS with the production schedule and Genetic Algorithm is an optimization approach used to improve the performance of the AS/RS operation by allocating the materials with minimum movement of shuttle. The effectiveness of Genetic Algorithm is demonstrated by applying it to the problem. The best sequences for parts with storage locations are obtained and conclusions are presented.

Keywords

AGV, AS/RS, FMS, Genetic Algorithm and Scheduling

Introduction

The increased demand for manufactured goods has increased the pressure on the manufacturing system, which in turn has motivated management to find new way to increase productivity, considering the scarce available resources. A Flexible Manufacturing System (FMS) has emerged as a viable alternative to the conventional manufacturing system. A FMS comprises automated machine tools, automated material handling and Automated Storage and Retrieval System (AS/RS) as essential components. Optimal performance of each element of this system would improve the productivity and efficiency of the total FMS. Existing FMS implementations have already demonstrated a number of benefits in terms of cost reductions, increased utilizations, reduced work-in-process levels, etc. However there are a number of problems faced during the life cycle of an FMS. These problems are classified in to: design, planning, scheduling and control problems. In particular, the scheduling task is very important because of the dynamic nature of the FMS such as flexible parts flow, Automated Guided Vehicle (AGV) routings and Automated Storage and Retrieval System (AS/RS) storage assignments. This work is primarily concerned with scheduling problems of FMS. This paper deals the scheduling of AS/RS integrated with parts scheduling in an FMS and gives near optimal solutions by using a non-traditional optimization technique, called Genetic Algorithm (GA).

Previous Research Work

Scheduling of FMS has been extensively investigated over the last three decades and it continues to attract the interest of both the academic and industrial sectors. Varieties of algorithms are employed to obtain optimal or near optimal schedules. Giffler and Thomson (1960) developed an enumerative procedure to generate all active schedules in a general 'n' job 'm' machine problem. Greenberg (1968) formulated a mixed integer programming (MIP) model for the 'n' job 'm' machine problem. Chan and Pak (1986) proposed two heuristic algorithms for solving the scheduling problem with the goal of minimizing the total cost of tardiness in a statically loaded FMS. Bigel and Davern (1990) showed the method of applying generic concepts to scheduling problems. He and Kusiak (1992) addressed three different industrial scheduling problems; heuristic procedures were developed to solve those problems. Kimemia and Gershwin (1985) reported on an optimization problem that optimizes the routing of the parts in an FMS with the objective of maximizing the flow while keeping the average in-process inventory below a fixed level. Umit Bilge and Gunduz Ulusoz (1995) proposed a time window approach to simultaneous scheduling machines and material handling system in an FMS. Gunduz Ulusoz et al (1997) proposed a genetic algorithm for the simultaneous scheduling of machines and material handling systems. Bozer and White (1984) developed expressions for the expected cycle time of an AS/RS performing single and dual command cycles. They used a continuous rack approximation to develop analytical models of the expected cycle time. Han et al (1987) improved the performance of an AS/RS by sequencing the order of servicing the retrieval requests. In this work, genetic algorithm is a non-traditional optimization approach used for scheduling both parts and AS/RS, simultaneously in an FMS.

Problem Description

Type and the operations of FMS differ with its configuration. Since the generalized configuration is not possible most of the research is confirmed to specific manufacturing systems. Configuration of the system along with the assumptions and the objective criteria considered in this work are presented in the following sections.

FMS Descriptions

The FMS (Jerald et al, 2005) considered in this work has the configuration as shown in figure 1. There are five Flexible Machining Cells (FMCs) each having two to six numbers of Computer Numerical machines (CNCs) each with an independent and self sufficient tool magazine, one Automatic Tool Changer (ATC) and one Automatic Pallet Changer (APC). Each cell is supported by one to three dedicated robots, for intra cell movement of materials between operations.

- There is a loading and unloading station from where parts are released in batches for manufacturing in the FMS and the finished parts are collected and conveyed to the finished storage.
- The five FMCs are connected by two numbers of identical AGVs. These AGVs perform the inter cell movements between the FMCs, the movement of finished product from any of the FMCs to the unloading station and the movement of semi finished products between the AS/RS and the FMCs.
- There is dedicated robot for loading and unloading of AGVs.

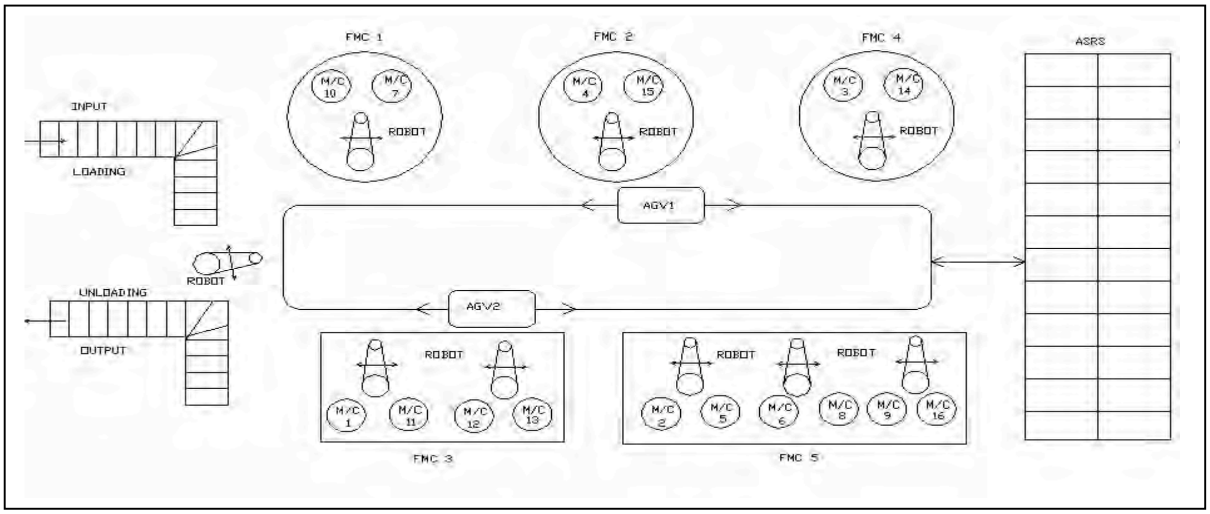


Figure 1. FMS Configuration

Model Assumptions

The Assumptions made in the work (Jawahar,1998) are

- There are 40 to 50 varieties of products for a particular combination of tools in the tool magazines
- Each type/variety has a particular processing sequence batch size, due date and penalty cost for not meeting the due date (Table 1).
- Each processing step has a processing time with a specific machine.

Part Number	Processing Sequence –{Machine number & Processing time in minutes)	Due date (Days)	Batch Size (Numbers)	Penalty cost (Rs/Units/Day)
1	{6,1},{7,1},{8,1},{10,2}	17	150	1.00
2	{2,1},{6,1},{8,1},{9,2},{14,4},{16,2}	17	200	1.00
3	{8,1},{11,3},{13,4}	14	800	1.00
4	{9,4}	26	700	2.00
5	{4,5},{5,3},{15,4}	11	150	1.00
6	{6,5},{14,1}	16	700	1.00
7	{3,5},{6,3},{16,5}	26	250	2.00
8	{5,4},{6,5},{8,1}	26	850	2.00
9	{4,1},{5,5},{8,1},{11,1}	1	100	0.00
10	{2,2},{9,1},{16,4}	20	150	2.00
11	{8,4},{12,2}	1	250	1.00
12	{6,2},{8,4},{10,1}	19	1000	3.00
13	{6,1},{7,5},{10,4}	25	700	4.00
14	{4,2},{5,3},{6,2},{15,2}	22	1000	4.00
15	{5,4},{8,3}	15	700	5.00
16	{5,3}	27	750	3.00
17	{3,1},{6,4},{14,1}	20	650	4.00
18	{9,2},{16,3}	24	250	5.00
19	{4,1},{5,5},{6,2},{8,2},{15,5}	5	450	1.00
20	{8,2},{11,4}	11	50	5.00
21	{4,5},{5,5},{6,2},{8,2},{15,5}	16	850	3.00
22	{12,5}	24	200	5.00
23	{4,2},{5,1},{6,5},{8,4}	14	50	4.00
24	{8,4},{11,4},{12,5},{13,4}	7	200	5.00
25	{7,3},{10,2}	24	350	1.00
26	{10,2}	27	450	0.00
27	{8,5},{11,5},{12,4}	22	400	1.00
28	{2,1},{8,1},{9,2}	3	950	5.00
29	{4,1},{5,5}	7	700	1.00
30	{11,3},{12,5}	18	1000	1.00
31	{8,2},{10,2}	2	800	2.00
32	{2,3},{6,4},{9,3}	15	800	1.00
33	{5,4},{6,5},{15,3}	27	500	4.00
34	{3,2},{6,2}	12	300	4.00
35	{3,4},{14,1}	9	900	2.00
36	{3,2}	20	700	2.00
37	{1,5},{2,2},{6,3},{8,3},{9,2},{16,4}	22	250	4.00
38	{2,4},{8,3},{9,2},{16,5}	8	50	1.00
39	{6,5},{10,5}	9	500	1.00
40	{2,2},{6,4},{9,4}	7	250	5.00.
41	{5,1},{8,2},{15,1}	22	800	4.00
42	{2,5},{6,4},{9,3},{16,1}	19	400	2.00
43	{1,3},{5,2},{6,2},{8,2},{15,3}	15	550	3.00

Table1: Machining Sequence, Time, Due Date, Batch size and Penalty Details.

AS/RS Structure

The structure of the AS/RS considered in this work as shown in figure 2. The storage structure considered is rectangular with 'm' rows and 'n' columns.

Each cell is capable of holding one item of any type. The center distance between any two adjacent cells in a row is Xr and in a column is Xc.

The numbers given inside the cell indicate the address of the storage location.

- There is one shuttle of the crane type, which is capable of moving vertically and horizontally.
- The P&D station is at the lower left hand corner of the aisle.

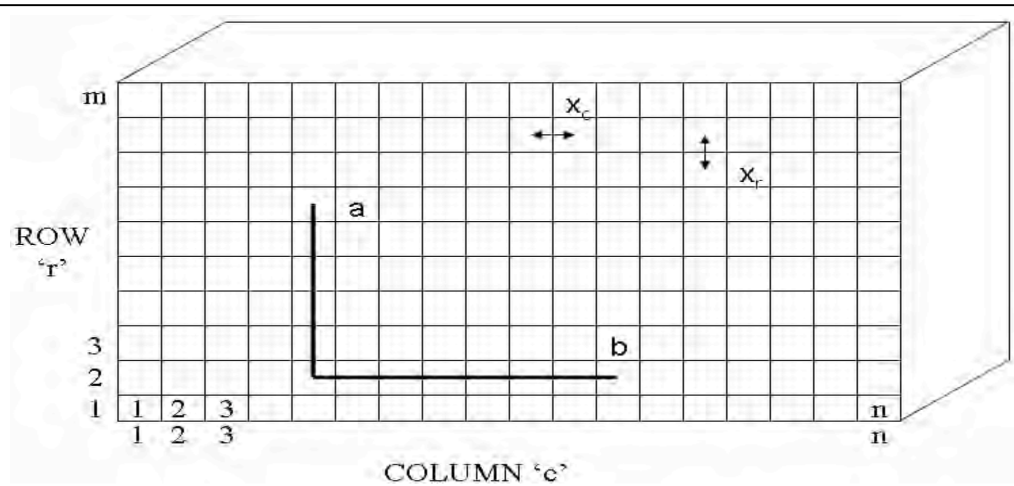


Figure 2. AS/RS Structure

In this AS/RS, Storage and Retrieval (SR) machine movement is a major decision parameter for its operation and control. The distance travelled by the S/R machine, when it moves from address a to b, to complete one activity is calculated by following two steps:

Step 1. Identification of row and column number

A simple algebraic equation can be used to ascertain the row and column numbers of the attributed address of the storage location. For example A is the address of the storage location its algebraic equation takes the form:

A = quotient * divider + remainder

Where divider is equal to number of columns n, then

$$\text{Row number } A_r = \begin{cases} \text{quotient} + 1, & \text{if remainder} \neq 0 \\ \text{quotient}, & \text{if remainder} = 0 \end{cases}$$

$$\text{Column number } Ac = \begin{cases} \text{remainder} & \text{if remainder} \neq 0 \\ \text{remainder} + 1, & \text{if remainder} = 0 \end{cases}$$

Step 2. Distance calculation for the movement from a to b

The distance between two storage locations a and b can be calculated using general formula,

$$P + X_r * (\text{mod } (a_r\text{-}b_c)) + X_c * (\text{mod } (a_c\text{-}b_c))$$

Objective Function

The objective is to minimize total distance travelled by S/R machine,

$$\textbf{Minimize } Z = \sum\nolimits_{i=1}^{l} P + X_r\ [\text{mod } (a_r\text{-}b_r)] + X_c\ [\text{mod } (a_c\text{-}b_c)]$$

Where,

Z – Total distance traveled by S/R machine

X_r & X_c – Centre distance between two adjacent cells in a row and column respectively

P – Distance from pick up and dispatch station to first storage cell

a_r, b_r – Distance between the locations a and b in row wise

a_c, b_c - Distance between the locations a and b in column wise

Genetic Algorithm

Genetic algorithm (GA) is stochastic search procedure for combinatorial optimization problems based on the mechanism of natural selection and natural genetics. These use the idea of survival of the fittest by progressively accepting better solutions to the problems. The algorithm starts with a randomly generated initial set of population called chromosomes that represent the solution of problem. These are evaluated for the fitness function and the selected according to their fitness value. The operators such crossover and mutation are works on it.

Genetic Operations

Reproduction

Rank order selection method is used for reproduction. The individuals in the population are ranked according to fitness, and the expected value of individual depends on its rank rather than on its absolute fitness. Each individual in the population is ranked in increasing order of fitness from 1 to N. After calculating the expected value of each rank, reproduction is performed using Monte Carlo simulation by employing random numbers.

Crossover

The strings in the mating pool formed after reproductions are used in the crossover operation. The possibility of cross over is determined by a probability between 0.4 to 0.9.A single point crossover is used for this work. If two strings are chosen for crossover, first a coin is flipped with a probability Pc = 0.6 to check whether a crossover is desired or not. If the outcome of the coin flipping is true, the crossover is performed; otherwise the strings are directly placed in the intermediate population for subsequent genetic operation. Flipping a coin with a probability 0.6 is simulated using Monte Carlo method. Crossover site is chosen by creating a random number. For example if the random number is 26, the strings are crossed after 26^{th} position. The current sequence after 26^{th} position the first string in that pair, is rearranged according to the next string. Similarly the sequence after the 26^{th} position of the 2^{nd} string in that pair is rearranged according to the first sequence.

Mutation

In this work, bit wise mutation is used. For mutation operation, two sites are selected by generating two numbers of random numbers. For example, if the random numbers generated are 14 and 36, then the corresponding job numbers in these positions are exchanged and the new sequence is obtained.

Basic Algorithm

Step 1: Generate random population of *n* chromosomes (suitable solutions for the problem)

Step 2: Evaluate the fitness *f(x)* of each chromosome *x* in the population

Step 3: Create a new population by repeating following steps until the new population is complete

Step 4: Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)

Step 5: With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.

Step 6: With a mutation probability mutate new offspring at each locus (position in chromosome).

Step 7: Place new offspring in a new population

Step 8: Use new generated population for a further run of algorithm

Step 9: If the end condition is satisfied, Stop, and returns the best solution in current population and Go to step 2

GA Parameters

The following genetic parameters are considered in this work.

Population size (n) = 20 samples
Crossover probability (Pc) = 0.6
Mutation probability (Pm) = 0.05
Termination criteria =100 generations

Results and Discussion

The designed scheduling procedure by Genetic Algorithm, software has been developed in the “C++” language. The best results obtained after performing 100 iterations, by using GA, are indicated in table 2. The parts schedule and AS/RS schedule for the given problem are obtained simultaneously by using the GA, for the best objective function value. The results shown that the GA is the best tool for producing near to optimal results for the scheduling problem of parts and AS/RS in the FMS. However, these results have not been tested with real environments because of unavailability of FMS easily. The method and suggestions can be incorporated in a real FMS to increase its productivity. The figure 3, shows the graphical representation of objective function values of GA.

Algorithm	**Parts Sequence**	**AS/RS Sequence**	**Objective Function Value**
Genetic Algorithm	27 3 15 24 18 23 17 33 9 6 21 20 28 2 38 10 34 41 11 37 32 31 16 22 12 8 26 7 42 30 19 36 4 5 14 25 43 1 35 29 40 13 39	17 10 31 16 36 47 13 19 2 33 18 35 5 46 43 6 27 8 48 37 20 9 3 30 39 22 34 4 40 45 29 49 28 25 24 26 1 11 15 23 9 15 20	4760 Units

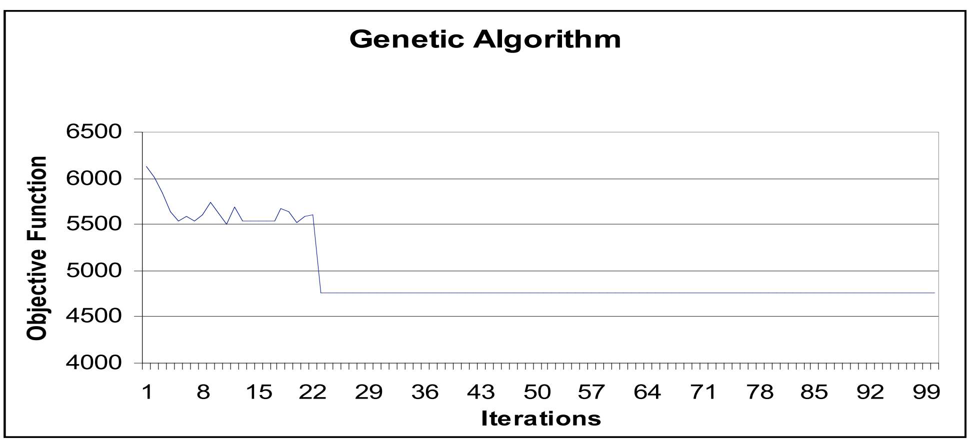


Figure 3. Graphical Representation of Results of Genetic Algorithm

Conclusion

In this paper, an attempt has been made to link the parts scheduling with AS/RS scheduling and the Genetic Algorithm, a optimization tool is used for getting the near to optimal results. The algorithm is used to find out the minimum movement of shuttle for optimum storage allocation of materials in AS/RS. The computational results have indicated that the proposed GA is very effective in generating near optimal solutions for simultaneous scheduling of parts and AS/RS. This type of approach can be used for various types of scheduling problems in the FMS.

References

Beigel., J. E., & Davern, J. J. (1990). Genetic algorithms and job shop scheduling. *International Journal on Computers and Industrial Engineering, 19(1),* 81-90.

Bozer, Y. A., & White, J. A. (1984). Travel time models for automated storage/retrieval systems. *IIE Transactions, 16(4),* 329-338.

Chan, T. S., & Pak, H. A. (1986). Heuristical job allocation in a flexible manufacturing system. *International Journal of Advanced Manufacturing Technology, 1(2),* 69-90.

Giffler, B., & Thomson, G. L. (1960). Algorithms for solving production scheduling problems *International Journal of Operations Research, 8*, 487-503.

Greenberg, H. H. (1968). A branch and bound solution to the general scheduling problem *International Journal of Operation Research, 16*, 353-361.

Gunduz, U., Funda, S. & Umit, B. (1997). A genetic algorithm approach to the simultaneous scheduling of machines and automated guided vehicles. *Computers and Operations Research, 24(4),* 335-351.

Han, M. H., McGinnis, L. F., Shieh, J. S., & White, J. A. (1987). On sequencing retrievals in an automated storage/retrieval system. *IIE Transactions, 19(1),* 56-66.

He, W., & Kusiak, A. (1992). Scheduling of manufacturing systems. *International Journal of Computers in Industry, 20,* 163-175.

Jawahar, N., Aravindan, P. & Ponnambalam. S. G. (1998). Optimal Random Storage Allocation for an AS/RS in an FMS. *International Journal of Advanced Manufacturing Technology, (14),* 116-132.

Jerald, J., Asokan, P., Saravanan, R., & Delphin Carolina Rani, A. (2005). Simultaneous Scheduling of parts and AGVs in an FMS using non-traditional optimization algorithms. *The International Journal of Applied Management and Technology, 3(1),* 305-315.

Kimemia. J., & Gershwin, S. B. (1985). Flow optimization in flexible manufacturing systems. *International Journal of Production Research, 23(1),* 81-96.

Umit, B. & Gunduz, U. (1995). A time window approach to simultaneous scheduling of machines and material handling system in an FMS. *Operations Research, 43(6),* 1058 1070.

OUTCOMES ASSESSMENT FOR UNDERGRADUATE MANAGEMENT DEGREE PROGRAMS: AN APPRECIATIVE INQUIRY APPROACH

Patricia Ann Castelli, Lawrence Technological University, USA
Richard Bush, Lawrence Technological University, USA
Virginia Kirkwood, Lawrence Technological University, USA

ABSTRACT

This paper describes one department's efforts to institute a collaborative and sustainable outcomes assessment process. The theoretical foundations of Appreciative Inquiry are provided and its applicability in creating an outcomes assessment plan for the undergraduate management programs is discussed.

Keywords

Appreciative Inquiry, education, organizational change, outcomes assessment, undergraduate management programs

Introduction to Outcomes Assessment

According to Castelli & Green (2006), assessing the outcome or result of any type of endeavor or event is not something that is capricious or new. We all conduct outcomes assessment on a regular basis informally as a professor, administrator, staff member, spouse and parent. Outcomes assessment is useful and necessary for our professional and personal success. We constantly evaluate our methods and results by asking: What task did I perform? What worked? What didn't work? What will I do the next time to improve the situation? And, how will I know if my plan was effective?

According to Rockman (2002),

> When the word assessment is mentioned, our reaction is often one of discomfort. Due to varying levels of understanding, we may associate assessment in terms of additional work in order to comply with external mandates from accreditation associations or governing bodies. Yet, if assessment is part of the annual planning, decision-making, and budgeting processes, there may be no additional work to perform. The data requested are likely to be the data we have (p. 1).

In higher education institutions, outcomes assessment *is* a critical part of gaining and maintaining regional and program accreditations. Like quality initiatives in business and industry, outcomes assessment assures quality in colleges and universities by evaluating the effectiveness of courses, programs, services, and the levels of student and staff satisfaction. Kretovics (1999) states that previous accreditation efforts focused on input measures such as teacher-student ratio, number of faculty with Ph.D.'s, number of volumes in the library and the breadth and depth of curricular offerings. But now the focus has shifted to value-added measures or student learning outcomes of the educational process.

Simply defined, outcomes assessment is a continuous process aimed at improving student learning outcomes and increasing an institution's overall effectiveness. An outcome is a final result or effect. According to Petkova & Jarmoszko (2006), the critical aspect of any assessment effort is the identification of learning goals (outcomes) for the program as a whole. Furthermore, Amin & Amin (2003) state that the curriculum and the instructional processes should help students achieve these objectives and outcomes. Outcomes are demonstrated as the competencies, knowledge and skills that a student is able to achieve or perform at the end of a course or program. Figure 1 illustrates examples of course and program outcomes.

Once outcomes are determined, means of assessment can be established. Based on the outcomes (competencies, knowledge and skills), faculty devise assessment methods to determine if students learned what they were supposed to learn in their courses/programs. Without assessment, learning outcomes cannot be measured.

Figure 1. Examples of Outcomes for Undergraduate Management Programs

Course Outcomes for a Business Major:

- Ability to apply effective decision-making for implementing financial and operating systems and controls
- Acquisition of skills to address ethical problems

Program Outcomes for a Business Major:

- Use of quantitative skills to effectively solve applied business problems
- Use of emerging technologies to appropriately gather, process, and present information

Course Outcomes for an Accounting Major:

- Application of accounting techniques and theories as they relate to business organizations
- Performance of information processing and transaction analysis for a service and merchandising concern

Program Outcomes for an Accounting Major:

- Prepare students for management positions in the accounting field
- Prepare students for the CPA examination

The Necessity of Outcomes Assessment

Colleges and universities engage in outcomes assessment for many reasons:

1. To ensure students are maximizing their learning experiences,
2. To ensure students are satisfied with the services and operational effectiveness of their college or university,
3. To provide effective techniques for identifying where changes and improvements are needed, and
4. To seek and maintain regional and program accreditations.

The benefits of outcomes assessment are numerous. Outcomes assessment increases institutional effectiveness by:

1. Providing opportunities for making fact-based changes and improvements
2. Effectively using data results to improve student learning outcomes
3. Increasing student satisfaction levels via changes and improvements to programs and services
4. Driving an institution's planning and budgeting processes
5. Strengthening the institution by demanding continuous improvement through ongoing assessment

Outcomes assessment can be summarized as a process of assessing and evaluating student learning, satisfaction levels, and operational effectiveness results to determine whether changes or improvements are needed. The outcomes assessment process that is developed for use by a particular institution should be one that the administration and faculty deem appropriate for their institution.

Steps in the Outcomes Assessment Process

The keys steps in the outcomes assessment process are shown in Figure 2. This paper will discuss the first step—creating an outcomes assessment plan document. Once an outcomes assessment plan document is created as described here, the next paper will discuss steps 2-5 in the process.

When devising an outcomes assessment plan, it is critical to involve faculty and staff early in this process. Without the contributions and support of a variety of members within a functioning unit, the process is likely to fail. Generally, an outcomes assessment coordinator is responsible for initiating, organizing, educating, and communicating the process of outcomes assessment. In most cases, the outcomes assessment coordinator is a full-time faculty member in the college he or she represents. Since the emphasis of outcomes assessment is on student learning outcomes, it makes sense to use a faculty member who is familiar with course and program functions. Moreover, in order to relate effectively with faculty and staff, it is critical to select or elect an individual who has a good rapport with the department and who serves as a 'cheerleader' in motivating faculty and staff to want to be a part of this process.

Figure 2. Steps in the Outcomes Assessment Process

1. Create the outcomes assessment plan document
2. Implement outcomes assessment measures per the plan document
3. Analyze and evaluate the results of assessment
4. Develop action plans for needed changes and improvements
5. Measure realized outcomes and report results

Since this process of outcomes assessment is relatively new to a functioning unit within an institution, a framework for initiating this process in a manner that involves the entire department was sought. Since the principles of appreciative inquiry support new and large initiatives such as outcomes assessment, integrating this approach was selected to ensure buy-in and commitment from faculty and staff.

Appreciative Inquiry Model

Understanding the foundations upon which appreciative inquiry is built enables a shift to a more positive change paradigm and intervention. "By embracing the constructionist principle organizations are free to create what they seek to create" (Bush & Korrapati, 2004). Appreciative Inquiry's (AI) strength rests in the unconditional positive question where the "notion that organizations are open books, which are continuously in the process of being co-authored and re-authored over time" (Ludema, Cooperrider, & Barrett, 2001, p. 189). The most critical step in deciding upon the focus of the inquiry is asking:

1. What topics should be studied?
2. What would the organization like to know more about?
3. What areas of discussion will potentially generate intriguing possibilities toward which the organization can work? (Cooperrider, Whitney & Stavros 2003; Ludema, Whitney, Mohr, & Griffin, 2003)

The topics can include most anything that affects or influences organization effectiveness. "Although one person or the management team of an organization can choose focus areas, topic generation is best accomplished through discussions with selected stakeholders at different levels of an organization" (Bush & Korrapati, 2004). Outcomes are very strong when different perspectives of the organization's current state and potential are integrated. The topics selected for inquiry will provide a starting point for development of formalized questions and will ultimately lead subsequent discussions and resulting growth in those directions (Bush & Korrapati, 2004). According to Cooperrider et al. (2003), "AI has demonstrated that human systems grow in the direction of their persistent inquiries, and this propensity is strongest and most sustainable when the means and ends of inquiry are positively correlated" (p. 30).

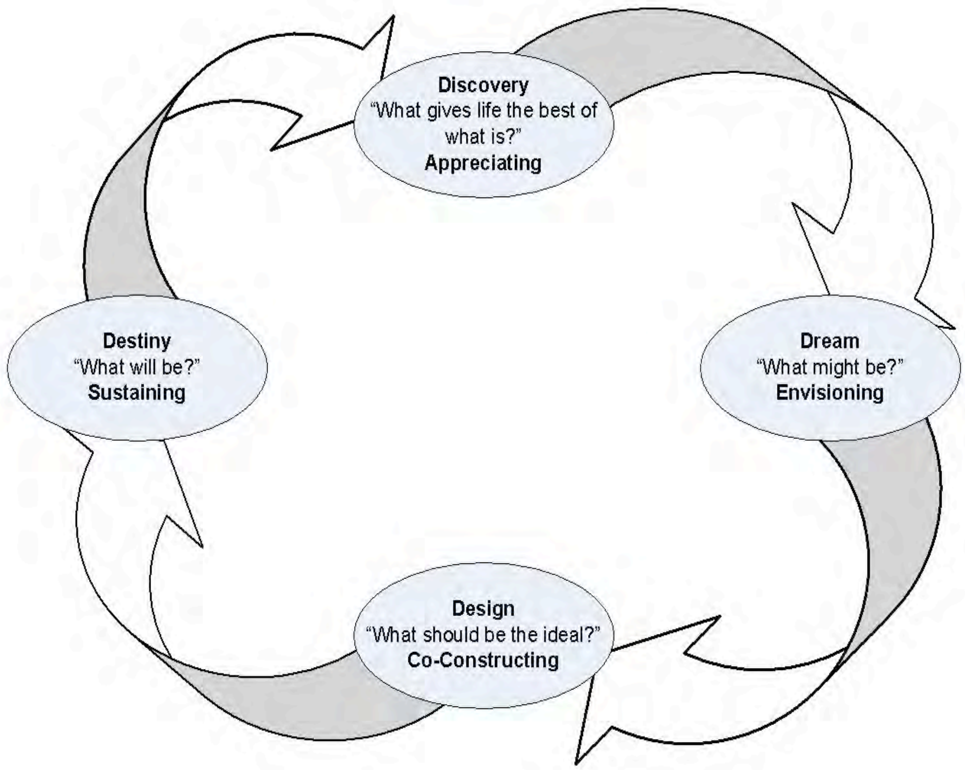
After the topic is identified, in this instance developing an outcomes assessment plan, questions are carefully crafted to elicit free-flowing discussion from interviewees regarding particular topic areas. In addition to being topic-focused, questions should be affirmatively constructed and posed in open-ended format whereby the participants are asked to "remember a time when… (Cooperrider et al., 2003)." Curran & Work (1998) describe it this way:

> As we were all familiar with traditional interviewing methods, we knew that for this nontraditional process, it would be essential to free interviewees from a state of mind of "answering" our questions; instead, we would encourage them to tell their own stories. The interview structure was meant to serve as a doorway into the person's meaningful recollections, not just to provide a means of gathering focused data (p. 247).

Effective questions and/or brainstorming should engage on both a personal and organizational level, thus defining a parallel between the reasons an individual has joined and continues with the group and clarifying the participant's social identity (Bush & Korrapati, 2004). Figure 3 illustrates the Appreciative Inquiry 4D Cycle.

Part of effective questioning is called institutional curiosity. According to Maki (2002), in the realm of outcomes assessment, institutional curiosity seeks answers to questions about which students learn, what they learn, how well they learn, when they learn, and explores how pedagogies and educational experiences develop and foster student learning. When institutions raise these questions and seek answers to them, assessment becomes a collective means whereby colleagues discover the fit between expectations for student achievement and patterns of actual student achievement.

<u>Figure 3</u>. Appreciative Inquiry "4D Cycle"



Source: Cooperrider, D., Whitney, D., & Stavros, J. (2003) Appreciative inquiry handbook: the first in a series of AI workbooks for leaders of change.

Discovery: What Gives Life? Participants are interviewed or work together in a brainstorming session in a process commonly referred to as the *discovery* phase—the first of a *4-D* cycle. The purpose is to find out what works in an organization, rather than to focus on what is ailing (Bush & Korrapati, 2004). This creates an opportunity to rediscover peak experiences, individual values, and strengths upon which to identify future opportunities and potential goals for the organization. Sessions may take place in large or small group situations or in one-on-one dialogs. Regardless of the forum, the interviewer should systematically ask prescribed questions, provide prompts and/or request clarification as needed, and withhold judgment and opinions that would sway interviewee responses (Cooperrider & Srivastva, 1987; Cooperrider, et al., 2003). The interviewer takes copious notes keyed to each question.

Dream: What Might Be? When possible, a larger group is convened for this second step in the AI process, the *dream* phase. In this phase, the results of the interviews are presented to the group, allowing the group an opportunity to identify themes or recurrent ideas revealed during the discovery phase (Cooperrider, et al., 2003; Whitney & Trosten-Bloom, 2003). The project leader asks questions to initiate discussion. Questions include:

1. What were the most humorous or surprising things you remember from the interview?
2. What one word best describes your feelings after the interview was completed?
3. What do you believe is the life-giving force of the Undergraduate Management degree programs? (Whitney & Trosten-Bloom, 2003).

A particularly effective way of helping to identify themes and simultaneously create enthusiasm for future visioning is to ask willing participants to recount some of the more compelling success stories (Bush & Korrapati, 2004). Common themes begin to emerge which speak to the organization's strengths and are used as the basis for collective discussion. Questions that further spur a group discussion include:

1. How can we capitalize on our strengths?
2. We have identified remarkable success when partnering with other organizations. How might we expand our pool of partners, and with which organizations do we think we would be most compatible? Why?
3. Where do we want to see the organization in five years? (Whitney & Trosten-Bloom, 2003).

"The dream phase is practical, in that it is grounded in the organization's history. It is also generative, in that it seeks to expand the organization's potential" (Cooperrider et al., 2003, p. 112). The dream phase provides an overall strategic focus for the organization.

Design: How Can it Be? The third *d* in the 4-D cycle stands for *design.* This phase builds on the dream phase and provides tangible results in the form of provocative propositions, sometimes called possibility statements, which are "an ideal state of circumstances that will foster the climate that creates the possibilities to do more of what works" (Hammond, 1998, p. 39). Provocative propositions challenge the status quo. They are goal statements which bridge the best of what is with what could be possible (Bush & Korrapati, 2004). Although challenging and inspiring, these hard and provocative propositions are attainable because they are developed from the program's successful past and identify its current strengths. These discussions should initiate action steps that translate intention into reality, what some have called creating the social architecture of the organization (Whitney & Trosten-Bloom, 2003). Through this design of an improved social architecture, the organization has the blueprint it needed so that leaders begin to realize the destiny that is within reach.

Destiny: What Will Be? The final phase of the AI process is the *destiny* phase. Based on a shared vision for the future, participants decide together upon a plan and commitments to implement the dreams that were refined in the design phase (Johnson & Leavitt, 2001). Cooperrider et al. (2003), provide the following description:

> It is a time of continuous learning, adjustment, and improvisation—all in the service of shared ideals. The momentum and potential for innovation are extremely high by this stage in the process. Because of the shared positive image of the future, everyone is invited to align his or her interactions in co-creating the future (p. 176).

While this is the final stage in the 4-D cycle, it represents a need to start the process from the beginning, creating an integrative approach to change anchored in an image of a positive future.

An Appreciative Inquiry Model for Outcomes Assessment

The first step in preparing for the process of developing an outcomes assessment plan involved a series of informal discussions between the College's outcomes assessment coordinator and both the director and assistant director of the bachelor of management programs. Conducting outcomes assessment was not new to our college of management, but was new for the recently acquired bachelor of management programs.

There are six main areas in an outcomes assessment plan. These areas include the mission and broad-based goals, student learning outcomes, basic skills development of students, personal development of students, accomplishment of goals and objectives, and planning and budgeting integration processes. Descriptions for each area are shown in Figure 4.

Using the framework of AI, the outcomes assessment team (consisting of the director and assistant director of the undergraduate management programs and the outcomes assessment coordinator for the college of management), brainstormed the initial focusing questions suggested by the AI process.

The following responses resulted.

1. What topics should be studied?

Figure 4. Key Content Areas of an Outcomes Assessment Plan

I. Mission and Broad-Based Goals

A. Provide the approved mission statement for the academic business unit.

B. State how you will provide evidence that the mission is being accomplished.

C. Provide the approved broad-based goals for the academic business unit.

D. State how you will provide evidence that the broad-based goals are being accomplished (Note: Some institutions use goals as strategies for accomplishing their mission).

II. Student Learning Outcomes

A. *Provide a summary of what students are expected to learn in a particular program (e.g., major or concentration). This is usually a summation of the learning objectives found in the syllabi of required courses.*

B. Identify a minimum of two direct and two indirect measures you will use to determine that the program expectations were met.

C. Explain how the direct and indirect measures will be used to determine that the program expectations were met (e.g., provide rubrics showing the criteria used to measure the results).

D. Copies of the instruments used as direct and indirect measures should be placed in the Appendix of the plan where appropriate.

III. Basic Skills Development

Prepare a statement on the "Basic Skills" development program at your institution and, in particular, how this is administered within your academic business unit.

IV. Personal Development

Prepare a statement on the "Personal Development" of students at your institution and, in particular, how this is administered within your academic business unit.

V. Accomplishment of Goals and Objectives

Explain how you will measure the effectiveness of your business unit (e.g., accomplishment of goals and objectives in the strategic plan is one of many different methods).

VI. Planning and Budgeting Integration Processes

Explain how the outcomes assessment activities will be integrated with the institution's planning and budgeting processes.

The team identified two major areas of study – implementing direct and indirect measures of outcomes assessment. Direct measures of outcomes assessment are defined as students' ability to demonstrate actual learning as a result of a course or program. Indirect measures rely on students' perceptions of the learning experience. Some of these measures existed prior to the plan while others needed development to aid in the creation of a more comprehensive plan.

The team decided to use the indirect measures of outcomes assessment already in place at the university and within the colleges. Utilizing the existing indirect measures across all four programs ensure consistency and, through the plan, coordinated ease of application for these measures. The measures include the Noel-Levitz Student Satisfaction Survey, the Graduate Survey and modification of the Mid-Term and End-of-Term Evaluations.

For direct measures of outcomes assessment, the team identified the International Computing Certification Professional (ICCP) exam for the information technology programs. For the management programs, a comprehensive exam for the capstone course Strategic Management was devised. This required the team, with the skills and knowledge of the professors who teach the course, to develop the exam and a corresponding evaluation rubric for its capstone course. Feedback from the team resulted in further revisions that were subsequently brought to all full-time faculty within the college of management for review. Discipline experts from each field added specific aspects to the exam including management and human resource content, finance and operational content, management information systems, and marketing management.

Once approved by the faculty who will be charged with evaluating the results, the exam was piloted by a group of 10 students enrolled in the Strategic Management capstone course. Revisions were made based on student and faculty feedback. Full implementation is scheduled for the fall term.

2. What would the organization like to know more about?

The office of undergraduate management programs seeks to learn more about the skills and abilities of its students and graduates. From the information provided by the direct and indirect measures of outcomes assessment, the team expects to identify best practices in ensuring course and program objectives are met, to provide innovative teaching and delivery specifically geared to the adult learner, to reduce redundancy between topics while expanding the depth of instruction on a given topic, and to build a positive environment where the results are applied to continuously improve the degree programs through better instruction.

The team leveraged lessons learned from previous outcomes assessment initiatives; faculty agreed that more emphasis should be placed on direct measures of outcomes assessment. The rationale is simple— direct measures provide concrete evidence of student learning in which students demonstrate acquisition of skills and knowledge as a result of learning during a course and/or program.

Administration of the various measures was also a factor in determining which value-added assessment items would provide the most meaningful results with minimal administration. This was due to the limited amount of support staff that currently exists in the department.

3. What areas of discussion will potentially generate intriguing possibilities toward which the organization can work?

The team, with input from faculty, determined that by addressing outcomes assessment, the programs can realize the potential of the organization as an open book, one that can be continuously re-authored (Ludema, Cooperrider, & Barrett, 2001). The larger topic area of outcomes assessment provides the starting point for the team to continuously develop and redevelop questions that lead to discussion and, ultimately, results that direct the team and programs in a positive direction for growth. The use of the results will aid the program offices to mentor and coach instructors using identified best practices and lessons learned, to minimize any overlap between subject courses, and to provide students with a quality positive experience.

The team decided further to incorporate technology into the outcomes assessment process as much as possible to ensure efficiency in gathering and processing information. In addition, redundancy assessments were eliminated and replaced and/or modified with instruments that are applicable to all programs.

Discovery: What Gives Life?

Recently four undergraduate programs were combined under one office to create the Undergraduate Management Programs. Three of the degree programs came by way of the humanities and engineering disciplines. The fourth existed in the management discipline. Two college deans oversee the administration and provide support for the four degree programs. Separately and combined, the four programs enjoy a strong reputation for quality which reflects the university's reputation for excellence. All of the management programs possess committed, knowledgeable, and professional instructors who are considered experts in their fields. The students are highly motivated, and many (91 percent) are working adult learners, have careers in their programs of study or are actively engaged in switching careers to management fields.

The topics identified by the team and faculty were:

- Combined Curriculum Quality
- Program/Combined Programs Reputation

The team identified outcomes assessment as the first and most crucial phase for addressing the topics since outcomes assessment is data driven and provides a strong basis for strategic planning. Furthermore, engaging in outcomes assessment is considered an important and vital activity to the university, for each college and for overall program quality. Also, by involving faculty throughout the process, the team was able to illicit their support thus gaining critical 'buy-in' for the implementation process. The team viewed these items as key success factors.

Dream: What Might Be?

The team adopted several provocative statements to create a design that would sustain outcomes assessment for the future. The statements include:

- To construct an outcomes assessment plan that supports and encourages an environment of continuous quality improvement throughout the curriculum as a tool for mentoring and coaching instructors, program and course reviews, and a measure of individual course performance and overall program quality.
- To construct a process for evaluating the data from the various instruments to identify and communicate changes to administrators and instructors in a timely manner, to use the results for constructive and proactive feedback, and to develop professional development workshops for faculty/adjuncts that address common themes that may emerge.
- To use the results of outcomes assessment to develop action plans aimed at increasing the effectiveness of courses and programs by making meaningful changes and improvements.

Important to addressing these statements is the collaboration between instructors, administrators, industry partners, and external accrediting bodies. The team and its partners will continue to connect the present to create a positive future.

Design: How Can it Be?

To address the provocative statements from the dream phase, the team constructed an outcomes assessment plan. The plan will be administered in a pilot phase over the summer semester, with revisions and a full implementation taking place in the fall. The team considers the Outcomes Assessment Plan a living document that provides a basis from which to constantly revise the plan to improve the outcomes and reporting necessary for accreditation, curriculum and instructor improvement, program growth, and the continued strengthening of the university's reputation for educational excellence.

Kretovics (1999) reinforces this approach quite elegantly:

> If learning is the ultimate goal of the educational experience, it is up to members of the institution to ensure this goal is met. Outcomes assessment is perhaps the best vehicle available at this time with great potential for affecting positive change and addressing the issues of accountability within higher education. The information obtained through an appropriately designed outcomes assessment program can be used by educators to continually improve the quality of academic programs offered and provide data and potentially answers to the concerns and questions of external and internal stakeholders (p.136).

Destiny: What Will Be?

The results of the Outcomes Assessment Plan will identify opportunities to:

- Improve instruction quality by identifying professional development needs, and mentoring and coaching opportunities;
- Improve student knowledge in all areas of their program of study;
- Develop a shared commitment and ownership of the OA plan by instructors and support staff; and
- Improve student retention by providing them with the skills and knowledge to apply for and receive admissions to a master degree program.

The team intends to identify new shared opportunities for change that positively affects student learning outcomes, and creates stronger skill-sets in our graduates with the active involvement of the program's instructor base. To accomplish this goal, faculty, staff and administrators of the Undergraduate Management Programs will meet monthly to share the results of the outcomes assessment with the entire team. Each semester the deans will be provided a consolidated report on outcomes assessment relative to each program. In addition, each year our advisory board members will be provided an annual report on outcomes assessment to elicit their input and comments on the activities and progress for the year.

Discussion and Further Research

The positive impact on morale, motivation, and results of the team's efforts contribute to addressing change with enthusiasm and energy. AI sets the stage for a winning outcome for organizational or program change initiatives. Throughout this planning process, the investigators used appreciative inquiry as a model for making positive and continuous change to reengineer existing undergraduate management programs. The results have been encouraging and energizing with success in advancing change while incorporating stakeholders throughout the university to affect that change. Further research includes continuing to evaluate the results of the implementation phase of outcomes assessment and action planning to ensure realized outcomes.

References

Amin, R.M., & N.A. Amin (2003). Benchmarking learning outcomes of undergraduate business education. *Benchmarking: An International Journal.* 10 (6). 538-558.

Bush, R., & Korrapati, R. (2004). *Appreciative inquiry as a model for positive change in an undergraduate information technology degree program.* Paper presented at the Allied Academies International Conference, Maui, Hawaii.

Castelli, P.A., & Green, J. Jr. (2006). *Outcomes assessment manual*. Prepared for the International Assembly for Collegiate Business Education (IACBE).

Cooperrider, D., & Srivastva, S. (1987). Appreciative inquiry in organizational life. In D. Cooperrider, D., Whitney, D., & Stavros, J. (2003). *Appreciative inquiry handbook: The first in a series of AI workbooks for leaders of change*. Bedford Heights, OH: Lakeshore Publishers.

Cooperrider, D., Whitney, D., & Stavros, J. (2003). *Appreciative inquiry handbook: the first in a series of AI workbooks for leaders of change*. Bedford Heights, OH: Lakeshore Publishers.

Curran, M., & Work, G. (1998). Creating opportunities for learning: A graduate class experiences AI in order to learn more about it. In S. Hammond & C. Royal (Eds.), *Lessons from the field: applying appreciative inquiry*: Practical Press, Inc.

Hammond, S. A. (1998). *The thin book of appreciative inquiry* (2nd ed.). Plano, TX: Thin Book Publishing Co.

Johnson, G., & Leavitt, W. (2001). Building on success: transforming organizations through appreciative inquiry. *Public Personnel Management,* 30(1), 129-136.

Kretovics, M.A. (1999). Assessing the MBA: what do our students learn? *Journal of Management Development* 18 (2). 125-136.

Ludema, J., Cooperrider, D., & Barrett, F. (2001). Appreciative inquiry: the power of the unconditional positive question. In P. Reason & H. Bradury (Eds.), *Handbook of Action Research: Participatory Inquiry and Practice*. Thousand Oaks: Sage.

Ludema, J. D., Whitney, D., Mohr, B., & Griffin, T. (2003). *The appreciative inquiry summit: a practitioner's guide for leading large-group change*. San Francisco, CA: Berrett-Koehler.

Maki, P. (2002). *Developing an assessment plan to learn about student learning*. www.aahe.org/assessment/assessmentplan.htm.

Petkova, O. & Jarmoszko, A.T. (2006). Assessment loop for the MIS program at Central Connecticut State University: A practice of learning, reflection and sharing. *Information Systems Education Journal* 4 (11) 3-9.

Rockman, I. F. (2002). The importance of assessment. *Reference Services Review*. 30 (3), 181-182.

Whitney, D. K., & Trosten-Bloom, A. (2003). *The power of appreciative inquiry: A practical guide to positive change* (1st ed.). San Francisco, CA: Berrett-Koehler.

GROUPING GENETIC ALGORITHM FOR CELL FORMATION WITH ALTERNATIVE ROUTINGS AND MULTIPLE MACHINES

K. Eswaraiah, National Institute of Technology-Warangal, India
C. S. P. Rao, National Institute of Technology-Warangal, India

ABSTRACT

In the present work authors have proposed a hybrid grouping genetic algorithm to solve cell formation problem in the presence of alternative routings for part types and replicate machines with an objective of minimization of inter cellular flow moves. This approach combines genetic algorithm with a local search heuristic for generation of optimal machine cells and part families. The grouping genetic algorithm is used to generate machine cells and part families and the local search heuristic is then applied on these machine cells to accelerate the process by reducing intercellular flow between cells. A Virtual Part Routing Matrix (VPRM) was proposed to represent the chromosome that avoids the duplicate genes and illegal crossover operation during reproduction. The developed approach eliminates the need of specifying number of cells and cell size during cell formation. Computational experience with the developed algorithm on a set of CMS problems available in the literature is presented. The approach produced solutions with a given objective that is at least as good as any results previously reported in literature and improved the objective function in many instances of the problems.

Keywords

Grouping Genetic Algorithm, Cellular Manufacturing System, Inter cellular Flow Movement, VPRM.

Introduction

Cellular Manufacturing System (CMS) have been extensively recognized as one of the most flexible and valuable approaches for companies switching from tradition practices to more lean and agile manufacturing techniques. CMS applies the concept of Group Technology (GT) in which similar parts are classified into part families and dissimilar machines are assigned into machine cells in order to exploit the cost-effectiveness of mass production and flexibility of job shop together can meeting the global competitive challenges. The design of CMS involves interrelated sub problems, namely machine grouping and part family formation (cell formation), intra-cell layout (machine layout) and inter-cell layout (cell layout) (Arvindh.B and Irani .S.A

1994). The process of determining part families and machine groups is referred to as the Cell Formation Problem (CFP) and which is the first and most important research topic in CMS, in the last three decades a large number of research papers and practical reports have been published in this area. Many solution methods such as mathematical programming, heuristics, optimization procedures and clustering techniques have been addressed to solve the CFP. In most cell formation methods parts and machines are assumed to have a unique part routing and machine types. However, it is well known that alternatives may exist in any level of a process plan. Explicit considerations of alternative process plans invoke changes in the composition of all manufacturing cells so that lower capital investment in machines, more independent manufacturing cells and higher machine utilization can be achieved (Hwang and Ree-96). A cell formation problem incorporating alternative process routings is called the generalized group technology problems. Relatively few research papers that deal with the cell formation problem with multiple identical machines and alternative part routes aiming at minimization of inter cellular flow movement. Kusiak (1987) was first described the cell formation problem in which alternative process routings were available. He presented an integer-programming model on the design of cells considering alternative process routings. Hark Hwang.L and Paek Ree (1996) was proposed a two-stage procedure for the cell formation problem with alternative part process plans for cell formation problem. At the first stage, process plans are selected for parts based on the compatibility coefficients while respecting the one part-one route principle. Two different ways of defining compatibility coefficients are suggested. At the second stage, part families are formed and machines are assigned. Proposed procedure is compared with Kusiak's generalized p-median model on four examples and results are found to be satisfactorily. Wom and Kim (1997) designed a generalized similarity coefficient and applied a multi criterion-clustering algorithm to obtain a machine cell. S. Sofianopoulou (1999) has been developed a model to solve cell formation problem under consideration of multiple copies of machines and alternative process plans for part types with an objective of minimization of the total intercellular moves. A two dimensional Simulated Annealing heuristic was employed to produce enhanced solutions to the problem of determining machine cells and part families. Yong Yin and Hazuhiko Yasuda (2002) developed a similarity coefficient that incorporates alternative routing, operation sequence, process time and quantity during the cell formation problem. They introduced five approaches to deal machine capacity violated issues during the cell formation Problem. Most of the CMS problems are solved by specifying number of cells and cell size during cell formation. Specifying such parameters is very difficult task in CMS design, some times, which leads to local optimal solutions in cell formation problems. To overcome such problems in this work a grouping genetic algorithm with local heuristic search methodology is proposed which eliminates the need for specifying number of cells and cell size during the cell formation. The remaining part of the paper is organized as follows: problem specification and formulation of objective functions is presented in Section 1, working mechanism and development of Grouping Genetic Algorithm for CFP is illustrated in Section 2, computational studies are reported in Section 3 and conclusions are given in Section 4.

Problem Specification and Formulation

In this paper, we attempt to solve the generalized group technology problem in CMS. Consider a manufacturing system having **'M'** machines and **'P'** parts. All parts (1,2..P) to be produced on a set of **M** machines (1,2…M). Each part is characterized by its operation sequence, process time, demand and batch size. Parts may have alternative routing and all parts are not having equal number of process plans. Each machine type can have multiple machines and these machine copies may be distributed in several different machine cells. Each machine is capable of doing a specific operation at a time and it is limited by a capacity constraint. Under a certain partition of machines for a part, there are several alternative part routings from entry, through machine cell(s) to exits i.e, each is associated with a different cost, in correspondence to the total expenses of inter cellular flow moves. It incurred when a part needs to travel from one cell to another cell in order to complete the part operations and it affects the material handling cost when the batches of parts have to be transferred between cells. Thus the problem is to find optimal machine cells and part families with their routings to minimize Inter Cellular Flow Moves (ICFM) between the cells. The mathematical model for cell formation problem, in the presence of multiple machines and alternative process plans is presented below based on the work of S. Sofianopoulou (1999).

$$ICFM = \min \sum_{p=1}^{p} \sum_{r=1}^{R_p} \sum_{s=1}^{S_p} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} C_{prsi} C_{pr(s+1)j} (1 - X_{ij}) Y_{pr} \qquad \ldots\ldots 1$$

Subject to

$$\sum_{r=1}^{R_p} Y_{pr} = 1; \quad p = 1, 2, ...P \qquad \ldots\ldots. 2$$

X_{ij}=0,1 i=1,…..,M-1,j=i+1,…,M. …….3

Where:

i,j 1...M machines
p 1...P parts
r 1...R_p alternative process plans for part type p
s 1...S_p operation sequence number for part type p
C_{prsi} a 0-1 coefficient indicating whether the s^{th} operation of the r^{th} alternative process plan of part p requires processing from machine i (C_{prsi}=1) or not (C_{prsi}=0)

Equation (1) calculates the total number of consecutive operations that are performed between any two machines not assigned in the same cell, for all machines and parts. X_{ij} is a 0-1 decision variable indicating whether machines i and j are assigned in the same cell (X_{ij}=1) or not (X_{ij}=0). Y_{pr} is 0-1 variable indicating whether part type p is processed following process plan r (Y_{pr}=1) or not(Y_{pr}=0).It ensure that all consecutive operations that are part of the same process plan for each part are taken into account.

Grouping Genetic Algorithm (GGA)

The introduction of the machine duplication as well as the presence of alternative routings for each part type into the cell formation problem, results in a model, that belongs to the class of quadratic 0-1 programs. Optimal solutions to these problems are computationally very expensive and only heuristic methods are capable of handling them efficiently. The heuristic that is proposed in this work is based on grouping genetic algorithm. Grouping genetic algorithm (GGA) is a group-oriented algorithm introduced by Emanuel Falkenauer (1998) designed to handle the special structure of grouping problems. Fakenauer developed GGA after realizing several significant drawbacks of using classical GA's for grouping problems and it is a special class of genetic algorithm, heavily modified to suit the structure of grouping problem. The crucial advantage of GGA is that it is able to deal with large instances of the problems, without the limitation of huge computational times. The concept of genetic algorithm is based on the evolution process that occurs in the natural biology. An initial population of individuals is generated randomly. In each generation, the individuals in the population are evaluated according to fitness function. For the next generation individuals are selected based on their fitness. New members of the population are generated by crossover and mutation operations. Reproduction focuses attention on high fitness individuals, thus exploiting the available fitness information. The process of reproduction, evolution and selection is repeated until a termination criterion is reached. The GGA varies from the standard GA in three ways: First the encoding scheme includes as augmented chromosome and groups are encoded on a one gene for one group basis. Second the Genetic operators work with the group section of the chromosome altering the composition of the groups. This in turn leads to alteration of the main chromosome. Third the GGA operators of crossover and mutation do not function in the same manner as performed in classical operators, in this crossover operator performs on the cell segment chromosome rather than machines and parts. The GGA presented in this paper uses a order based chromosome encode and an evolution strategy identical to one proposed by Evelyn .C .Brown and Robert.T.Sumichrast (2001).The advantage of order based chromosome representation is that all offspring's formed by crossover and mutation are feasible solution and perhaps natural representation of a tour. The hybrid GGA approach presented in this paper combines a grouping genetic algorithm with a local search heuristic. The genetic algorithm is used to generate sets of machine cells and part families. The evolutionary process, embedded in the genetic algorithm is responsible for improving the grouping quality of the sets of machine cells and part families. The local search heuristic is applied to the sets of machine cells generated by the genetic algorithm to improve the objective function. As mentioned earlier, the objective function used is the minimizations of inter cellular flow moves. The other important aspects of genetic algorithms: chromosome representation and decoding, parent selection, crossover and mutation will be discussed in the following sections.

Chromosome Encoding

In this work, order based four-segment chromosome representation is developed. The advantage of order based chromosome representation in cell formation always results in a feasible solution, since the constraint for the number of machines in each cell does not exist within the genetic structure. A chromosome encoding for cell formation problem is shown below:

$$C_1\ C_2 \ldots C_G \mid M_1\ M_2 \ldots M_M \mid P_1\ P_2 \ldots P_P \mid R_1\ R_2\ R_3 \ldots R_R$$

Where

C_G = Number of cells in the solution c=1 to C_G

M_M = Number of machines in the system m=1 to M_M

P_P = Number of parts in the system p=1 to P_P
R_R = Number of routes of each parts between 1 to r_R

The first segment holds the numbers of cells, the second segment to holds the machines sequence, third segment holds the part sequence and forth represents the routing information of each part. Each chromosome is a vector sum of c+ m + p + r. For a given machine component incidence matrix (MCIM) the number of machines, parts are constant, where as the number of cells vary in length in each chromosome. Each part may have different routes. To incorporate the multiple machines in the chromosomes a special encoding structure is developed. In this it is assumed that the total number of machines is equal to number of unique machines plus similar machines. The similar machines are given numbers in the increasing order after the unique machines. For example a four unique machine string of 1 2 5 4 3 6, in this the machine 5 and 6 are duplicate of machine 2 and 4 respectively. The developed program has the capability to extract the features of the similar machines during the calculation of objective function. The proposed scheme eliminates the illegal gene (machine) structure during initialization of population and crossover operation.

Encoding of Part Routing

The initial population generation of part routing sequence and crossover operation may lead some times illegal gene structure because all parts are not having equal number of process plans. A typical problem may arise in the part routing segment of genetic operations while seeking solutions for the said problem. The problem is explained in the following two steps. For illustration consider a problem size of 4x6 and number of process plans available for each part is shown in Table 1.The chromosome encoding scheme are described in the following steps.

Part Number	No. of Routings
1	1
2	4
3	3
4	2
5	2
6	2

Table 1: Part Routing Data

Step 1: **Select two chromosomes from the population to represent part sequence and routing portion. For example,**

	Part	Routes
Parent1:	1 2 3 4 5 6|	1 4 3 1 2 2
Parent2:	4 5 2 3 1 6|	1 2 2 1 1 2

Step 2: Perform two point crossover operations on part and routing segment of the chromosomes and the corresponding offspring's are yields as:

Part Routes

Offspring1: 5 2 3 4 1 6 | 1 2 2 1 2 2

Offspring2: 4 1 2 3 5 6 | 1 4 3 1 1 2

In offspring 1, part number 1 is assigned with a route number 2, where as it is having only one part routing in the part routing data, hence it is violating the input condition. To overcome such difficulties a Virtual Part Routing Matrix (VPMR) heuristic is proposed and its methodology is illustrated in following steps.

Step1: Read the part routing information and find the maximum number of process plans (MPR) of each part.

Step2: Generate NPXMR (Number of Parts x Number of Routes) route matrix for all parts as shown in Table 2.

Part Number	No. of routes			
	1	2	3	4
1	1	1	1	1
2	1	2	3	4
3	1	2	3	1
4	1	2	1	2
5	1	2	1	2
6	1	2	1	2

Table 2: Virtual Part Routing Matrix

For example, part 1 having only one routing; the developed VPRM generates a four-routings equal to original process plan. The Part 3 as having three routes whereas maximum numbers of process routes are four. For this case it is not possible to give the equal probability of routings, to find the fourth routing a random number generated between 1 and 3 and assigned it as the fourth route i.e. route 1. Part four having 2 routes, then it generates route 3 and route 4 equal to route1 and route 2 respectively. Similarly part 5 and part 6 also. In the similar fashion, process will be continued until completion of all part routings. Now the number of routes is equal for all parts, i.e., four. The proposed method eliminates the need of repair function during the initial population and crossover operation.

Evaluation of Fitness Function

Each chromosome in the population is evaluated for a given objective function. In the present work inter cellular flow moves (ICFM) is considered as the objective function, which is a minimization function. Generally GA's will find a chromosome with maximum fitness value for better solution. To do this the objective function is maximized by using F(i) = -F(i), where F(i) is the fitness value of each individual.

Reproduction Policy

A mixed selection strategy based on binary tournament selection and elitist selection is adopted. It combines the features of random and deterministic sampling for selection of chromosomes for next generation. In this two strings are selected randomly from the population and compared; the string with higher fitness is selected to participate in reproduction.

Selection for Crossover Operation

In this work, authors considered a pool management strategy for selection of chromosomes for crossover and mutation operations. For this 20% pop size of best strings in the current population are copied in to the next generation population called as elitism. The rest of the new population is created by the usual genetic operations applied on the entire current population. This process provides that the best solution of the current population is not only passed from one generation to the other but also participates in the reproduction process.

GGA Crossover Operator

The primary purpose of the crossover operator is to permit new regions in the search space. In the present work-grouping crossover operator is implemented and in this some portions of cells of each parent chromosome are exchanged to form offspring (E.C. Brown et al 2001). The crossover operation is explained by considering a cellular manufacturing system having 10 machines and 12 parts. Each part having maximum of four routes only and crossover methodology illustrated by following steps.

Step1.Select randomly two parents from the population, say parent1 and parent2 respectively as shown below:

Parent1: 1 3 2| 8 2 4| 6 1 5 10 |9 3 7 | 11 1 4 3| 8 2 6 |10 9 5 7 12| 1 2 3 1 1 2 3 2 1 4 2 1
Parent2: 1 3 2 4| 10 6 |1 4| 3 2 8 |7 5 9 | 9 12 1 |3 7 2| 4 8 10 11 |5 6 | 2 1 4 2 2 1 3 3 1 2 1 4

Step2: Assign machine, part and route to corresponding cells. Schema of chromosome representation for parent1 and parent are shown in figure 1 and figure 2 respectively.

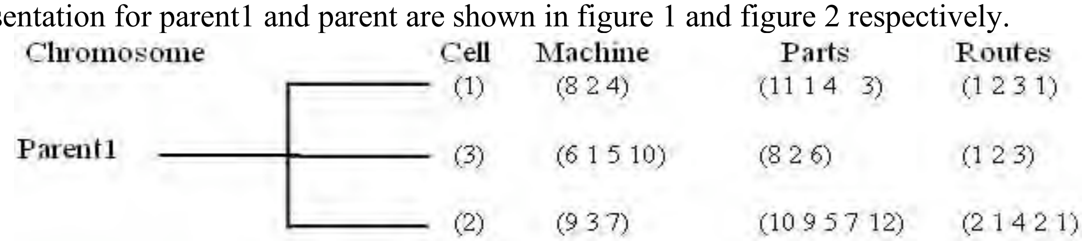


Figure 1: Schema of Chromosome-Parent 1

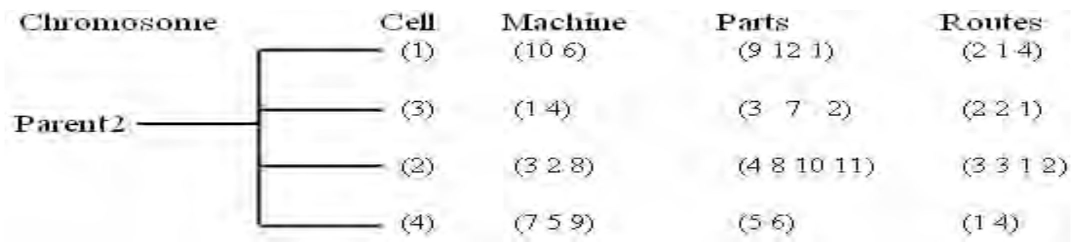
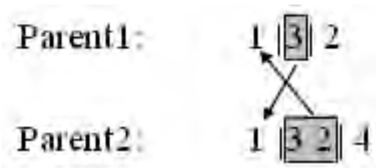


Figure 2: Schema of chromosome-Parent 2

Step 3.Generate two random numbers between 1 to Nc-1 and cut the cell group section of parent 1 and repeat the process for parent 2. The developed program ensures that it does not allow for two cut points on the same location of a chromosome and the selected crossing section for parents are shown below:



Step 4: Inject the contents of the crossing section of the first parent at the first crossing site of the second parent and vice versa as shown below:

Offspring1: 1 3 2 3 2

Offspring2: 1 3 3 2 4

Now offspring1 and offspring2 are the two new chromosomes and bold number represents the injected cells.

Step 5: The allocation of machine, part and route for offspring 1 is shown in figure 3.

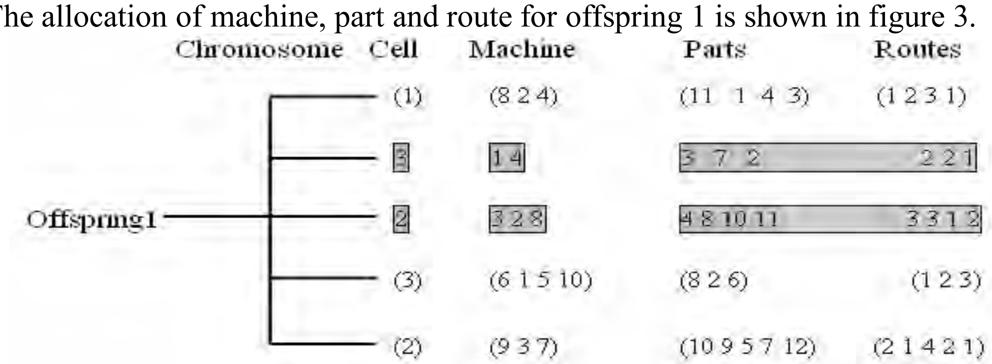


Figure 3: Schema of chromosome-offspring 1

Step 6: The formed off spring1 may have illegal gene structure because of duplicate genes in the machine and part groups. The result process may be infeasible some times like parts assigned to zero machine groups and vice versa. To overcome these difficulties the following heuristic method is employed to assign the parts and machines to cell groups.

Step 7: Assume that the injected section group remains intact and the other groups are subject to alternation. In this example **3(1 4) (3 7 2) and 2(3 2 8) (4 8 10 11)** remains same in offspring1 and other groups are subject for alteration. A repair algorithm is applied for elimination of duplicate machines and parts. In this example cell one formed with zero machines and assigned one part family which is an infeasible solution and to avoid such situation the following step is used.

Step 8: For displaced machine, examine the existing groups that resulted from crossover and determine which of them contains the most machines from the machine flow matrix and place displaced machine in that group and repeat Step 8 for all displaced machine and same logic can be extended to parts also. After applying the heuristic procedure offspring1 yields as:

Offspring1: 3 2 1 4| 1 4 3 2 8 6 5 10 9 7 | 3 7 2 4 8 10 11 6 1 9 5 12|2 2 1 3 3 1 2 3 2 1 4 1

From the crossover operation it was observed that even though parent1 consists of three cells the resulting offspring1 would have four cells. This indicates that a cell group segment is variable in length and cell sequence changes in each crossover operation needs keeping the injected cell numbers intact and other cell numbers are chosen randomly to facilitate reduction of intercellular moves. The total GGA works on this principle to produce optimum number of cells and their corresponding objective function.

Mutation

In this work a swap mutation operator is employed with a probability of 0.1, for this two random numbers are generated between 1 and number of machines or parts, and then genes in between the selected positions are reversed within the group segment of machine or part. For example in offsprin1 gene position 2 and 6 in the machine group is selected and corresponding machine values are interchanged. After mutation final offspring1 shown as:

Offspring1: 3 2 1 4| 1 4 3 2 8 6 5 10 9 7 |3 7 2 4 8 10 11 6 1 9 5 12|2 2 1 3 3 1 2 3 2 1 4 1

Offspring1: 3 2 1 4| 1 6 3 2 8 4 5 10 9 7 |3 7 2 4 8 10 11 6 1 9 5 12|2 2 1 3 3 1 2 3 2 1 4 1

The whole process of evaluation, selection for crossover and mutation is repeated for number of generations. The entire procedure for calculation of objective function is represented in the form flow chart and shown figure 4.

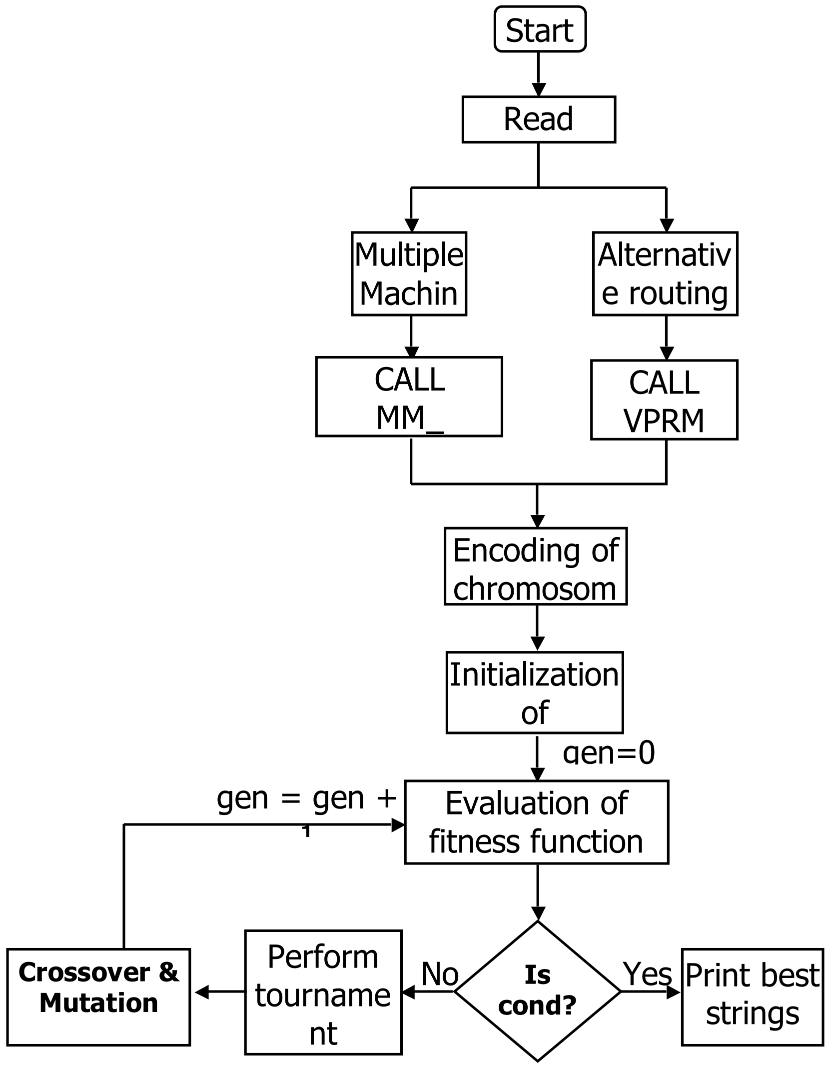


Figure 4: GGA flow chart for ICFM

Computational Studies

Relatively few research papers that deal with the cell formation problem with multiple functionally identical machines and alternative process routes aiming at minimization of ICFM in CMS. To demonstrate the performance of the proposed algorithm, we tested the hybrid grouping genetic algorithm on CMS instances collected from the literature varies from 4x5 to 20x30 (Machines × Parts). Mostly cited CMS data sets are binary 0-1 matrixes which does not have part routing sequence and quantity. For these problems it was assumed as part operation sequence in ascending order of the machine number and quantity are taken as unity and the corresponding inter cellular flow moves are computed. The algorithm was coded in visual C^{++} and we use a PC with Pentium (1.5GHz) processor to test our implementation. The present state of the art practice on genetic algorithm does not provide information on how to configure them.

Therefore a small pilot study was conducted in order to obtain a reasonable configuration. The algorithm was configured as follows and the configuration was held same for all problems. The number of chromosomes in the population equals three times the number of machines in the problem. The crossover probability varied from 0.4 to 0.8 on test problems. In most of the cases the good solutions are obtained for a probability of 0.6 for small size problems (upto 10 machines) and a probability of 0.8 observed for large size problems (upto 20 machines). In all the cases mutation probability found as 0.1.It is observed that as the mutation probability increases the objective value decrease in a generation due to separation of good machine groups in a cell. The probability during the crossover made equal to 0.8 and mutation rate taken as 0.1.The elitist strategy copies to the next generation best 20%of the current population. Each run of the algorithm was replicated 25 times and for the above parameters the average number of generations needed to reach the best solution was calculated. For all the problems the developed hybridGGA method seeks less than 150 generations to find the best solution.

The developed hybridGGA method applied on 10 data sets and our proposed method outperforms over best-known values (published in the literature) on three of the 10 data sets, and matches its performance on the remaining problems. Both the descriptions on the test problems and the final computational results on 10 test problems are shown in Table 3. In order to demonstrate the effectiveness of the proposed model and methodology in solving cell formation, first five problems are taken from the work of S. Sofianopoulou (1999). He presented a cell formation under alternative process planes and multiple machines with an objective of minimization of ICFM using simulated annealing algorithm. The proposed algorithm was applied on these problems; our approach yielded the solutions with the fewest intercellular flow moves (ICFM) for the 3, 4 and 5 problems. The maximum percentage of objective function improvement found in these problems is 12%. With respect to problems 1 and problem 2 the hybridGGA methodology yielded the same results. The improvement of objective function over best-known value for test problems is shown in figure 5 and optimal machine cells and part families for test problems are presented in Table 4.The optimal machine component incidence matrix(MCIM) for the problems 3, 4 and problem 5 are shown in figure 6, figure 7 and figure 8 respectively**.** In the same manner our methodology obtained same results for problems 6,7,8,9 and 10.The solutions generated by the our hybridGGA method using ICFM objective function was analyzed to gain an understanding of why they were superior to solutions obtained by other methods (best known values). One is the developed method eliminates the need for specifying number of cells and cell size well before the solution of the cell formation problem. Specifying number of cells and cell size may restrict the solution to be trapped in the local minima and thus the proposed method overcomes trapping in local minima and seek global minimum. Second one is the proposed VPRM heuristic generates efficient part routings and it eliminates the illegal gene mappings during grouping crossover, further reduces computational time and appeared to be fast converging.

P_No	Author	MxP	NC	CS	Multiple Machines	Alternative Routing	ICFM		INCFM
							LOFV	AOFV	
1	S.Sofianopoulou 1999(P1)	4x5	2	2	No	Yes	0	0	4
2	S.Sofianopoulou 1999(P2)	12x20	3	5	No	Yes	29	29	36
3	S.Sofianopoulou 1999(P3)	12x20	3	5	No	No	31	28*	36
4	S.Sofianopoulou 1999(P3.1)	12x20	3	6	2(2),5(2)	Yes	25	22*	39
5	S.Sofianopoulou 1999(P4)	16x30	3	7	3(2),7(2)	Yes	34	32*	54
6	Yong Yin &Kazuhiko Yasuda 2002(P1)	6x8	4	4	No	Yes	10	10	15
7	Yong Yin &Kazuhiko Yasuda 2002(P2)	5x7	3	3	No	Yes	0	0	11
8	Chao-Ton Su & Chih-Ming Hsu 1996(P1)	18x30	6	6	15(2),17(2) 18(2)	No	100	100	1815
9	Nagi.R et al 1990	20x20	5	5	No	Yes	1	1	46
10	Sankaran.S & Kasilingam.R 1990	6x10	2	4	No	Yes	2	2	16

Table 3: Problem Description and Computational Results

NC = Number of cells; CS=Cell size; LOFV=Literature objective function value, AOFV=Author objective function value, INCFM=Intracellular flow moves

P_No	Source	P_size	*NC*	Machine Groups	Part groups Part no(route no)
1	S. Sofianopoulu Prob.1	4x5	2	1 3 4 2	2-2,4-2,5-2 1-2,3-2
2	Prob.2	12x20	3	4 1 6 11 5 12 8 2 3 9 10 7	2 5 7 9 10 14(2) 16 19 1 4 6 8 11 13 17 3 12 15 18 20
3	Prob.3	12x20	3	1 3 10 4 2 12 9 5 11 6 8 7	1 13 15 17 2-4,10 14 18 19 5-9 11 12 20
4	Prob.3.1	12x20	3	5 6 11 2 3 7 12 5 8 2 1 4 9 10	2(2) 5(2) 6(2) 11(2) 16(2) 17(2) 1(2) 13(2) 20 9 7 18 15(4) 3(3) 4 8 10 12(3) 14 19
5	Prob.4	12x30	3	10 3 4 9 16 6 12 7 11 15 1 14 2 13 3 5 7 8	7 27 20 3(3) 14 10 28 23 22 19 21 18 15 4 5 8 16 1(2) 13 29 12(3) 30 25 17 9 24 11 2(2) 26 6(3)
6	Yong Yin & Yasuda.K (2002) Prob.1	6x8	2	1 4 2 3 5 6	1(1) 4 8 6 3 2 5(2) 7(2)
7	Yong Yin &Yasuda.K (2002) Prob.2	6x8	2	1 4 2 5 3	1 2 5 3 4(2) 6 7
8	Chao – Ton Su & Chih-Ming Hsu (1996)	18x30	4	4 15 11 14 8 1 14 13 18 10 3 17 2 18 15 7 12 6 16 17 5 9	2 4 7 9 11-13 20 23 29 18 25 28 30 1 3 5 6 16 19 21 22 24 26 27 8 10 14 15 17
9	Nagi.R et al 1990	20x20	5	3 14 12 9 13 17 16 4 5 1 15 19 6 20 11 8 7 2 10 18	17 18 19 13-16 5 6(6) 7(2) 8(6) 1(2) 20(2) 2(2) 4(3) 3(2) 10(2) 11(2) 12(2)
10	Sankaran.S & Kasilingam.R 1990	6x10	2	4 5 2 6 1 3	1 2(2) 4 6(2) 8 9 10(2) 7 5(4) 3

Table 4: Optimal machine groups and part families for test problems

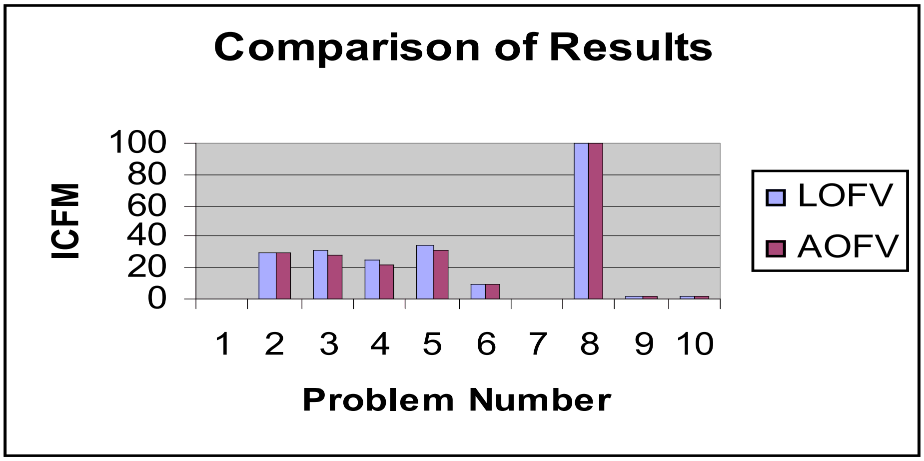


Figure 5: Improvement of Objective Function

MACHINES \ PARTS	3	4	7	8	9	10	14	15	19	20	1	2	6	11	12	13	16	17	18	5
1	4	2				1												2		
3		4			3		1		1		3			2			4	1		2
2	2			2	1	4		2	2	3			3	3			5			
10	1	3					3		3			1			2	1				
7					2	2	4			2		5				2	2			
4	3	1		3		3	2	3										4	1	
6		5								1	1	3	4	6			1			3
5	5							1			2	4	2	5	3	4			3	
8			2								5		1	4	4			3	2	
11	6				4						6	2			1	3	3			
9				1		5			4					1						
12			1		5				5		4									1

Figure 6: Machine Component Incidence Matrix for Problem 3

MACHINES	P	2	5	6	11	16	17	1	9	13	20	7	18	15	3	4	8	10	12	14	19
	R	2	2	2	2	2	2	2	1	2	1	1	1	1	3	1	1	1	1	1	1
	5	4						2		4					5						
	6	3	3	4		1	4	1			1										
	11	2			4	3	3		4	3					6				1		
	2		2	3	3	5	1		1		3										2
	3				2	4	2	3	3							4				1	
	7	5				2			2		2							2		4	
	12		1		1			4	5	2		1									5
	5			2	6								3	1					3		
	8			1	5			5				2	2						4		
	2													2	2		2	4			
	1														4	2		1			
	4														3	1	3	3		2	
	9																1	5			4
	10	1								1					1	3			2	3	3

Figure 7: Machine Component Incidence Matrix for Problem 4.

MACHINES	P	16	14	3	19	22	21	20	7	10	28	1	23	6	13	26	8	27	18	4	29	9	15	25	11	30	2	12	17	5	24
	R	1	1	3	1	1	1	1	1	1	1	2	1	3	1	1	1	1	1	1	1	1	1	1	1	1	2	3	1	1	1
	10	1	1	1	4	1				3	3		2	2			1		4										4		
	3			3			2					1	3		3																
	4			2	3	3	1	1	1	4	2						4	1										1			
	9	3		4								2		4	2				3								4		1		
	16	4	2		1				4		1																				
	6	2				2	4	3		5				3		2				2					4						
	12		3		2	4			2				1			3	2	3						1					2		
	7									1						1				5							3				
	11							2	5								3				2	1	2			3				1	
	15						3					4			1				1				3	2	1		2	3			
	1													1								4			2						
	14												4							4			1		3	2	1		3	2	
	2										4	3	5					2	5	3		5		4		1					
	13									2										1		3		3				2			
	17(3)																														
	5								3												3									3	2
	18(7)																				1										3
	8																		2			2								4	

Figure 8: Machine Component Incidence Matrix for Problem 5

Conclusion

In the present work a hybrid grouping genetic algorithm has been developed to solve the cell formation problem for manufacturing systems where multiple functionally identical machines exist and alternative process plans for part type available. The objective of the model is to minimization of inter cellular flow movement. The processing sequences for each part type is taken into account in order to determine the exact amount of inter cellular flow moves between the cells. Hybrid grouping genetic algorithm approach was employed to produce efficient solutions to the problem of determining machine cells and part families and process routings. A virtual part routing method was proposed to eliminate the repair function during crossover and mutation operation and reduces the number of iterations during the execution of the program. The developed model is tested using the well-known problems from the literature. In most of the previous investigations stipulate that the total number of manufacturing cells should be known in advance and the upper and lower bounds of the number of machines in each manufacturing cell should be specified for solving the machine cell group problem. The developed algorithm eliminates the need for specifying number of cells and cell size. It is inferred from the results obtained in all the data sets are executed at the specified number of cells and cell size constraints (as specified in the literature) the developed algorithm showed equal results and in some cases improved results with the literature data.

References

Aravindh. B., and Irani, S.A. (1994). CF: The need for an integrated solution of the sub problem. *International Journal of Production Research*, pp.1197.

Caux. C., Bruniaux, R and Pierreval, H. (2000). Cell formation with alternative process plans and machine capacity constraints: a new combined approach. *International Journal of Production Research*, 64, 279-284.

Chao–Ton Su and Chih-Ming Hsu. (1996). A two phased GA for the cell formation problem. *International Journal of Industrial Engineering 3(2)*, 114-125.

Evelyn C. Brown, and Robert. T. Sumichrast (2001). CF-GGA: a grouping genetic algorithm for the cell formation problem, *International Journal of Production Research,* Vol.39, No.16, 3651-3669.

Falkenauer E. (1994). A new representation and operators for GAs applied to grouping problems. *Evolutionary Computation*, Vol. 2, No. 2, pp. 123-144.

Goldberg, D.E. (1989). Genetic Algorithms in Search Optimization, and Machine Learning. *Addison-Wesley*, Reading, MA

Harhalakis, G, Nagi, R and Proth, J.M. (1990). An efficient heuristic in manufacturing cell formation for group technology applications. *International Journal Production Research 28,* 185-198.

Hwang .H and Ree. P. (1996). Routes selection for the cell formation problem with alternative part process plans. *Comput. Ind. Engg.*, 30, 423-431.

Jun Wang. (2003). Formation of machine cells and part families in cellular manufacturing systems using a linear assignment algorithm. *Automatica* 39, 1607–1615.

Kazerooni, M, Luong, L H S and Abhary, K. (1996). Cell formation using genetic algorithms. *Int. J. Flexible Automation and Integrated Manufacturing,* Vol.3, No.3-4, 219-235.

Kusiak, A. (1987). The generalized group technology concept. *International Journal of Production Research,* 25(4) 561-569.

Nagi, R, Harhalakis. G and Proth. J. M.,(1990). Multiple routings and capacity considerations in group technology applications. *International Journal of Production Research*. 28, 2243-2257.

Rajamani. D, Singh. N and Aneja. P. (1990). Integrated design of cellular manufacturing systems in the presence of alternative process plans, *International Journal Production Research 28,* 1541-1554.

Sofianopoulou.S. (1999). Manufacturing cells design with alternative process plans and/or replicate machines. *International Journal of Production Research*, vol.37, No.3, 707-720.

Sankaran.S and Kasilingam.R.(1990). An integrated approach to cell formation and part routing in group technology manufacturing system. *Engineering Optimization* 16, 235-245.

Srinivasan. G and Narendran T. T. (1991). GRAFICS-a nonhierarchical clustering algorithm for group technology, *International Journal Production Research*, *29, 46-78*.

Yong Yin and Hazuhiko Yasuda (2000). Manufacturing cells design in consideration of various production factors. *International Journal of Production Research*, Vol.40, 885-906.

Zhao.C.W and Wo, Z.M. (2000). A GA for manufacturing cell formation with multiple routes and multiple objectives” *International Journal Production Research*, Vol.38 No. 385-395.

A Collective Airport-Airline Efficiency Strategic Model

Reza G. Hamzaee, Missouri Western State University, USA
Bijan Vasigh, Embry-Riddle Aeronautical University, USA

ABSTRACT

A new approach to airline sustainability is proposed here by adoption of an internalization of the existing externalities between airlines and airports through an interactive financial strategy. It is hypothesized in this paper that airports could optimize their revenues from airlines' operations by reducing their fees to such optimal levels as those necessary to induce more flights, and therefore, more passenger movements into and from airports. Secondary (landside) airport businesses would be exposed to higher demand by the increasing number of passengers, through which airports' other sources of income will be more secured than otherwise. The concept of Laffer curve (Blanchard, 2000, p.453) is applied here to some optimal airport taxation of airlines. This would enhance airport efficiency, which could facilitate airport fee reductions even in the short run. In an ideal special case, our model provides for a practical integration of airports and commercial airlines into one unified enterprise (say, "airportlines"), in which the main output would be a two-tier (and complementary) output. Using this model, one may apply the step-wise regression along with linear programming procedures, to explore optimum airport fees and allocation of resources through a collective cost minimization and/or revenue maximization, including those of the airlines, airports, and an integrated "airportlines."

Keywords
Airlines, Airports, Laffer Curve.

Introduction

What the airline industry is experiencing today is a serious structural problem along with several adverse external factors, such as fuel, increasing costs of anti-terrorism, etc. Airline-airport partnership and alliances are more critically needed than before. The airlines management, under a variety of unfriendly business conditions, is a challenging task in this day and age. From fuel prices to government regulations and still harsh competitive environment, airlines should find their innovative ways of survival.[1]

1. The merger between Bankrupt US Airways and America West may provide a better opportunity to compete more effectively with discount rivals and complement each other geographically.

On August 17, 2005, the price of jet fuel stood at $1.87 per gallon, the highest quarterly price ever recorded. Total fuel expense accounted for 19.2% of the industry's total operating costs, a 4.5 percentage increase over year. The massive increase in the price of fuel caused a 30.4% increase in unit fuel costs to 2.23 cents per Available Seat Mile (ASM).[1]

The recent Delta and Northwest Airlines' bankruptcy filings are the latest consequences of the troubled industry.[2] United Airlines, US Airways and ATA are operating under Chapter 11 bankruptcy protection, and other carriers appear to be hovering on the edge.[3] Delta Airlines, the nation's largest carrier lost $2.2 billion in the fourth quarter of 2004, a total of $5.2 billion loss for the year due to the cost of jet fuel. Four of the six largest carriers are now reorganizing their finances under court supervision. Since 1978, there have been more than 100 airline bankruptcies, including UAL Corp.'s United Airlines (2002), the biggest ever.[4] Commercial airports around the world are also reducing their dependence on government support and make their operations more effective. In U.S., commercial airports have tried to control their costs and improve services by collaborating with the private sector.[5]

One of the possible solutions is the merging of each airport with the airlines utilizing that airport. In this study, the authors are providing a merger model of operation, in which airports and airlines have common goal of promoting the collective prosperity of their own, while incorporating the welfare of the passengers and customers into their operation model. In some alternative business models, considering the fact that airport incomes from landside and airside businesses, would heavily depend on the airlines' success, it is recommended that the fees charged by airports could be reduced in such tough times as the currently experiencing extremely high-gas prices. The overall incomes of the airports would be higher, hypothesizing that the airlines demand for usage of airports is an elastic demand with respect to various airport fees.

Airports and airlines collaboration is crucial for promoting mutual survival and profitability. Fore example, airport retail sales in Western Europe reached to about $6 billion in 2002, up 13% from the previous year. In Japan, Osaka's Kansi Airport offers a wide selection of restaurants and shops. Brussels Charleroi Airport, a small airport with less than 200,000 passengers per year, responded positively to Ryanair proposal to forgo the airport landing fees. Consequently, Ryanair now flies about two million passengers a year. In 2002, Ryanair serviced 12 destinations from Charleroi, generating a traffic of 2 million passengers a year.[6] In U.S., the airport retail volume is estimated to be $3.5 billion a year, and is ranked third in airport concession revenue behind airport parking and car rental businesses. This revenue could be generated only if airlines are successful in bring more passengers to the airport.

Some Background Studies

According to the Federal Aviation Administration (FAA, 2002) there are 546 commercial airports. Of these, 422 have more than 10,000 enplanements and are grouped as commercial service airports in the United States. Furthermore, there were 33 large hubs, with 464,486,847 enplanements, 35 medium hubs, with 115,177,169 enplanements, 68 small hubs, with

1 . For more information see, the Air Transport Association of America (ATA, 2005).
2 . Both Delta and Northwest Airlines filed for Chapter 11 on Wednesday, September 14, 2005.
3 . United Airlines' loss for the 4th Quarter was $664 million, which totaled to $1.6 billion for the 2004 entire year.
4 . The Wall Street Journal, September 19, 2005, A1.
5. Combining Private Equity, Economic Development and Transportation, Overview of Transportation Public-Private Partnership Project Financing, Lehman Brothers, December 16, 2003.
6. Moreover, Ryanair operation in Charleroi led to the creation of 200 direct jobs.

50,202,980 enplanements, and 410 non-hub airports (faa.gov).[1] Only a few hubs have pure capacity problems, the main hubs generally are dominated by one of major carriers, which makes entry by another airline difficult (Button, 2002). At most commercial airports, the financial and operational relationships between an airport and the airlines it serves are defined in legally binding agreements, called "airport use agreements", that specify how the risks and responsibilities of airport operation are to be shared between the two parties. These contracts specify the terms and the conditions governing the airlines' use of an airport and delineate the rules for calculation of the compensation airlines must pay for use of airport facilities and services, and to identify the airlines' rights and privileges. The airport airline financial relationship at the nation's major commercial airports is based on the following approaches with important ramifications for airport pricing and investment practices:

- The Residual Cost Approach: under which the airlines collectively assume significant financial risk by agreeing to pay any costs of running the airport that are not allocated to other users or covered by non-airline sources of revenue (FAA 1997). So, the airlines agree to keep the airport financially self-supporting by making up any deficit remaining after the costs identified for all airport users have been offset by non-airline sources of revenue.[2]

- The Compensatory Approach: under which the airport operator assumes the major financial risk of running the airport and charges the airlines fees and rental rates set so as to recover the actual costs of the facilities and services that they use.[3]

- The Hybrid Approach: some airports have adopted a hybrid approach, combining both residual and compensatory methodologies. For example, the agreements at the Washington airports have revenue-sharing elements. At the end of each year, the profit or loss from all the airport operations including commercial facilities is split between the airport authority and the airlines. The profits assigned to the airlines will be put towards off-setting the next year's fees.

Analogous to other enterprises, it is customary for airports to publish their accounts at yearly intervals. The exact format of the published annual financial report varies significantly among airports within individual countries. Often the form of accounting is determined by the form of ownership, which itself varies significantly among different countries. Total revenue of an airport is frequently divided into two principal categories: operating revenue and non-operating revenue.

Operating revenues are revenues that are directly associated with the running and operation of the airport, including the operational areas, terminals, leased areas, and grounds. Examples of such revenues are landing fees, terminal area concessions, and airline-leased areas. Most U.S. airports use non-aeronautical revenues to offset total airport expenses before setting

---

1. FAA classifies airports based on the percentage of national total passengers enplaned. According to FAA, a large hub airport is defined as the one with 1% or more of national enplaned passengers. These percentages for medium and small hub airports are 0.25 to 0.99% and 0.05 to 0.24%, respectively (EEEA 2003).

2 . Airport System Development (Washington, D. C.: U.S. Congress, Office of Technology Assessment, OTA-STI-231, August 1984).
3 . Ibid.

the landing fees. These airports charge as much as the market could bear for the concessionary operations. This cross-subsidization may steer to over-utilization of the airfield facilities. Practically, in many countries airports use aircraft weight as the basis to price their airfield facilities and services.[1]

Non-operating revenues, on the other hand, are all income to the airport operator derived from activities not directly associated with the operation of the airport. These revenues would presumably continue even if the airport were to shut down. Examples of such revenues are contributions from government, interest earned on airport capital funds, and revenues received from sales of property and plant.

Just as with revenue, airport expenditures can be classified as both operating and non-operating. The operating expenditures of an airport are those expenditures incurred in the course of running the airport. These expenditures can disappear if the airport operation is closed down. These expenditures are typically grouped under:

- Traffic handling and commercial activities.
- Salaries, administration expenses, and transport expenses.

Airport non-operating expenditures are those expenses incurred even if no operations are carried out. They may be grouped according to:

- Interest payments on outstanding capital.
- Amortization on fixed assets, fees for various non-operating purposes, and the like.

In most major U.S. hub airports, there are both airside and ground side capacity problems.[2] The advent of hubbing or hub premium has intensified the concentration of arrivals and departures unevenly during the day. The mix of long and short haul and international/domestic mix will affect the airport's peak and off peak characteristics. In order to respond to the above problems or manage them, the airports should either expand their infrastructure or utilize the demand management technique as an immediate solution.

Airline investment in airport infrastructure is not new. In August 1996, two Chinese airlines became the first to gain permission to build support facilities for their aircraft and passengers at the new Pudong International Airport.[3] In April 1997, British Airways and the Port Authority of New York and New Jersey, invested $250 million at JFK's terminal 7, which included an additional building expansion with new retail outlets and concessions. The project also included new access roadways and improved parking facilities. In 1999, American Airlines started construction of a new $1.4 billion terminal complex to replace terminal 9 and eventually terminal 8 at JFK. In June 2003, the new terminal 2 at Munich airport opened. According to the airport and Lufthansa it is the first time an airline has entered into joint entrepreneurial responsibility for financing, building and operating an airport. Under the agreement, the airport takes on a 60% stake in terminal 2, with Lufthansa holding the 40% balance. Lufthansa

1. Nonetheless, the weight-based approach is not effective for congested and over-utilized airports, it may be appropriate for airports that are underutilized. Peak hour pricing for the airfield facilities may alleviate the congestion at many over-utilized airports.

2 . Airside capacity problems such as; takeoff and landing slots, enplaning and deplaning gates are more noticeable problems in those airports that have one single dominant client. Atlanta airport is dominated by Delta so Minneapolis/St. Paul is dominated by Northwest Airlines.

3. The China Business Information Network, August 30, 2006.

contributes knowledge of traffic development and hub control coordination, functional terminal planning and IT passenger information systems (Johnson 2001).

In August 2003, an Indonesia-based airport service company formed an agreement with Singapore Airlines (SIA) to create a maintenance services joint venture.[1] Under the agreement, Singapore Airlines Engineering Company (SIAEC) takes a 49% share in the new company while JAS holds the remaining 51%. Lufthansa expanded its presence in the booming South China market by setting up and international airfreight terminal at Shenzhen airport, which will be able to handle 200,000 tons of cargo annually. Shenzhen Airlines may take a controlling stake in the expected joint venture, while Lufthansa will mainly contribute its operational and management skills in 2004.

A Theoretical Model of Airline-Airport Integration

The authors propose the following theoretical framework for an airline-airport integrated operation optimization, in which all three stakeholders, the airlines, airports, and their customers (of both airside and landside services) are incorporated. Various solutions for group optimization would be analyzed.

Beginning with the two general groups of airside and landside outputs to be produced, there are n different resources to be used by both airlines and airports. Therefore, the n resource constraints are defined as:

$$\begin{aligned} a_{1AS} \bullet Q_1 + a_{1LS} \bullet Q_2 &\leq R_1 \\ a_{2AS} \bullet Q_1 + a_{2LS} \bullet Q_2 &\leq R_2 \\ \cdots \\ a_{nAS} \bullet Q_1 + a_{nLS} \bullet Q_2 &\leq R_n \end{aligned} \qquad (1)$$

where,

a_{iAS} = the amount of the ith resource necessary to produce one indexed unit of airside output (landing & departure), for i = 1,2,…, n

a_{iLS} = the amount of the ith resource necessary to produce one indexed unit of landside services to customers at the airport, for i = 1,2,…, n

Q_1 = the total indexed quantity of airside output (quantity of a composite output of landing/passengers + take off/passengers + miles/passengers, or alike)

Q_2 = the total indexed quantity of landside output

R_i = The total quantity of the ith utilized resource, for i = 1,2,…, n

In Figure 1, as an illustrating example, five of the aforementioned hypothetical resource constraints are graphed. Obviously, to arrive at a relevant production possibility frontier (the darker portions of the five constraints), all of the n resource constraints listed in above In (1) must be simultaneously implemented. That would be essentially the integrated airline-airport production possibility frontier. This model is proposed to include only one airport (at a time) as integrated with all the airlines chartered to have movement (traffic) through it. An integration of all the resource constraints for one airport and all the airlines using that airport would be summarized in constraint (2), as follows next:

1 . PT Jasa Angkasa Semesta (JAS).

$$\sum_{i=1}^{n} (a_{iFL} \bullet Q_1 + a_{iLS} \bullet Q_2) \leq \sum_{i=1}^{n} R_i \qquad (2)$$

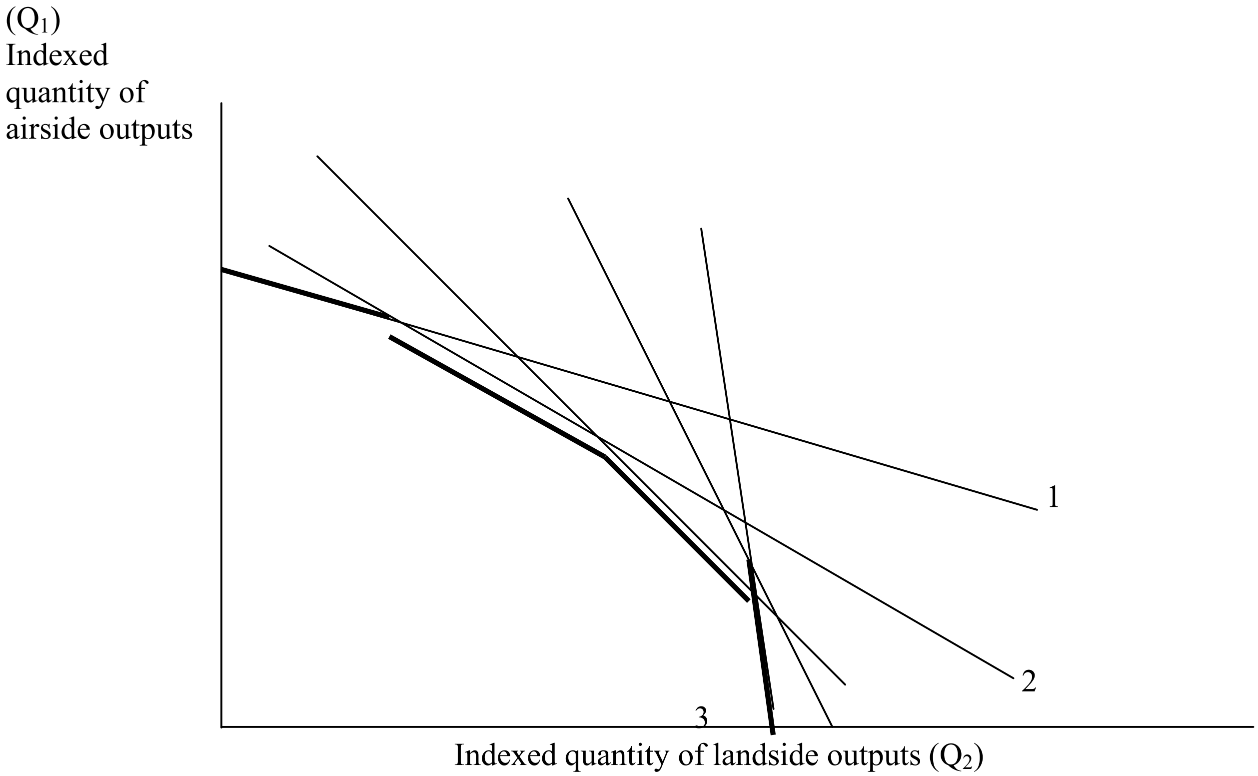


Figure1
Airport-Airline Resource Constraints

That is an integrative resource constraint of one airport - only – along with those of all the airlines using it. Such resources, as an example, could include - but not limited to - the following list:

R_1= quantity of Gas
R_2= number of pilots
R_3= number of airside personnel
R_4= number of aircraft
R_5= number of runways
R_6= number of maintenance bases
R_7= number of maintenance technicians and engineers
R_8= number of tower controllers
R_9= amount of computer hardware and software to utilize
R_8= number of airlines' on-land employees excluding airside personnel, technicians and engineers

R_7= number of landside operational employees
R_{10}= number of security personnel
R_{11}= number of janitorial employees
R_{12}= number of value of security facilities
R_{13}= number of airport restaurants

The Airlines-Airport Budget Line

Then under competitive conditions, the following condition should hold:

$$P_1 = ATC_1 \qquad (3)$$

Also, by definition,

$$ATC_1 = \frac{\sum_{i=1}^{k} r_i R_i}{Q_1} \qquad (4)$$

which is the average resource cost, considering k different resources to be used for provision of airside output (Q_1).

P_1 = Indexed average price of a composite unit of output (landings/passengers, departures/passengers, plus miles/passengers of traveling)

Q_1 = Quantity of a composite output of landing/passengers (q_1) + takeoff/passengers (q_2) + Miles/passengers (q_3)

ATC_1 = average total cost of all resources needed for each composite unit of the airside output produced in a certain period of time

r_i = the rental price (cost) of the ith resource in production of airside output for i = 1,2, …, k

Comparing (3) and (4), the following definition, under competition, will result:

$$P_1 = \frac{\sum_{i=1}^{k} r_i R_i}{Q_1} \qquad (5)$$

Also, P_2, the price of an indexed quantity of landside output (Q_2), can be similarly defined as:

$$P_2 = ATC_2 = \frac{\sum_{j=1}^{m} r_j R_j}{Q_2} \qquad (6)$$

where:

ATC_2 = average total cost of all land-side output supplied in a certain period of time

r_j = the rental price (cost) of the jth resource in production of landside output for j = 1,2, …, m

n = k + m

Then the following relationship (7) will represent the budget constraint for the passengers and/or general customers, which would also represent the airlines-airport budget constraint, assuming that their incomes under competitive conditions would be the same as their total costs:

$$P_1Q_1 + P_2Q_2 \leq B$$
$$B = B_1 + B_2 \qquad (7)$$

B_1 = Airlines total budget

B_2 = Airport's total budget

$B = B_1 + B_2$ = Airportlines' total budget

Now, plugging (5) and (6) in (7), the following budget constraint will be resulted:

$$\frac{\sum_{i=1}^{k} r_i R_i}{Q_1} \bullet Q_1 + \frac{\sum_{j=1}^{m} r_j R_j}{Q_2} \bullet Q_2 \leq B$$

$$Q_1 = \left(\frac{B}{\frac{\sum_{i=1}^{k} r_i R_i}{Q_1}} \right) - \left(\frac{\frac{\sum_{j=1}^{m} r_j R_j}{Q_2}}{\frac{\sum r_i R_i}{Q_1}} \right) Q_2 \qquad (8)$$

(Q1-Intercept) (Slope of the Budget Constraint)

Q_1 = (Total airport-airlines integrated budget) / (Airlines average cost of operation)

$$- \left(\frac{ATC_{landside}}{ATC_{airside}} \right) \cdot Q_2 \qquad (9)$$

Now all the three stakeholders are put together into interaction, and find various possible optimization solutions to the model. In the following figures, Q is a product transformation curve (production possibilities frontier), which is simply defined as the locus of all the possible two-output combinations that can be (efficiently) produced by full utilization of all the available resources. The product transformation curve for landside and airside outputs is shown in Figures 2,3, and 4. Obviously, the rate of product transformation (RPT) is defined as the quantity of one product or service that must be forgone in order to obtain one more unit of the other without varying the currently employed input level. (Henderson and Quandt, 1976).

In the following figure, consumers of both services would have a different optimal solution than the "airportlines" would. The consumers' preferences are more heavily towards airside than land-side services. However, for the "airportlines" more of the landside and less of the airside would be the best solution.

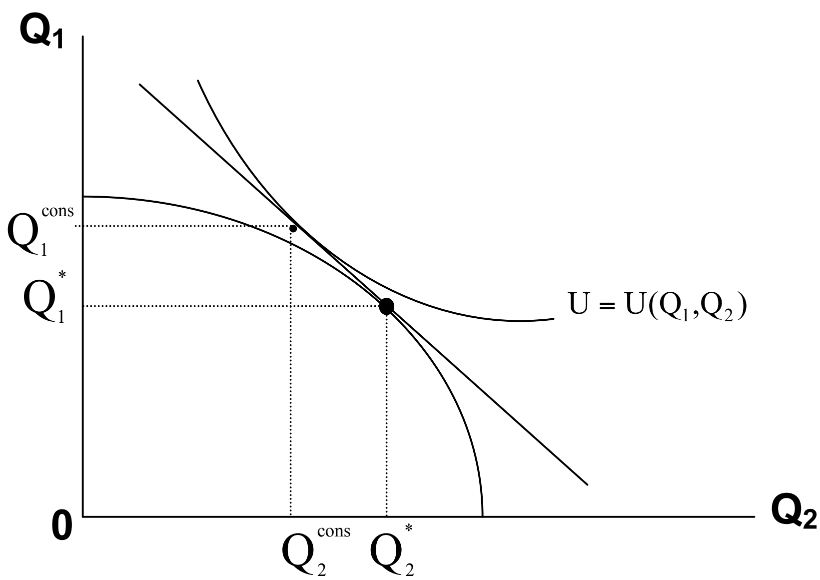


Figure 2
Consumers of both services would have a different optimal solution than the "airportlines" would produce.

In the following case (Figure 3), the providers (airports and airlines or just "airportlines") will have again a different optimal solution than the consumers would. Consumers reveal to prefer more landside than airside services. The determination of the optimum production strategy must take into consideration, in addition to the relative prices of landside and airside services, the cost of inputs to the system. In the long run the initial and final costs of building a new runway, terminal buildings, passenger concession area and other regulatory expenses must additionally be taken into account in a prorated fashion, as applied to one unit of output provided per unit of time.

The major determinants of profitability in the aviation industry are the cost and the prices received for providing landside and airside services. With changing technological and marketing opportunities and constraints, current costs and prices must be viewed as temporary.

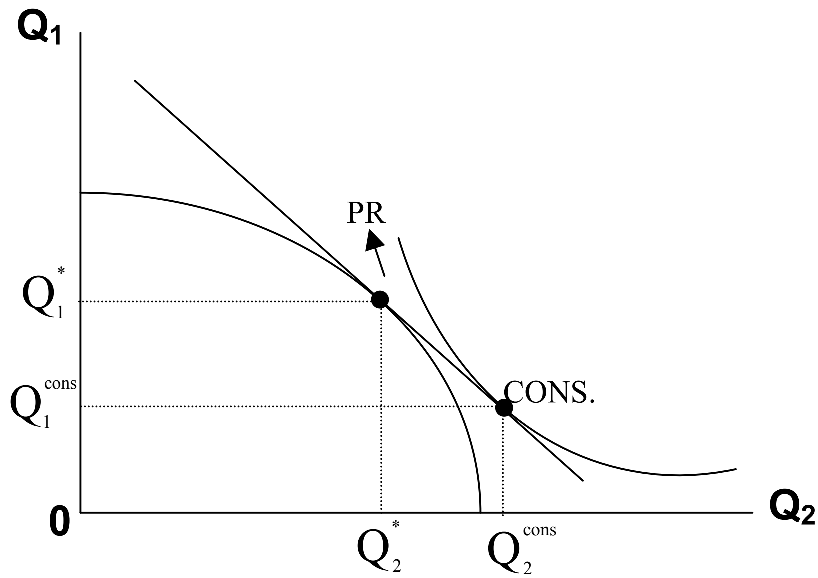


Figure 3
Consumers' biases are towards landside services

In Figure 4, the optimum solution for all three groups is the same, Q^*_1 and Q^*_2 should be produced and consumed to bring all the parties up to their highest satisfaction and efficiency levels altogether. While this last theoretical solution is less practical and more hypothetical, like all other abstract theories, it could be used to approach to what is most practically and collectively advantageous to the social welfare level as well as the specific parties involved. Among other things, the entire model can be used to analyze the optimum investment in infrastructure both at the airside and landside facilities.

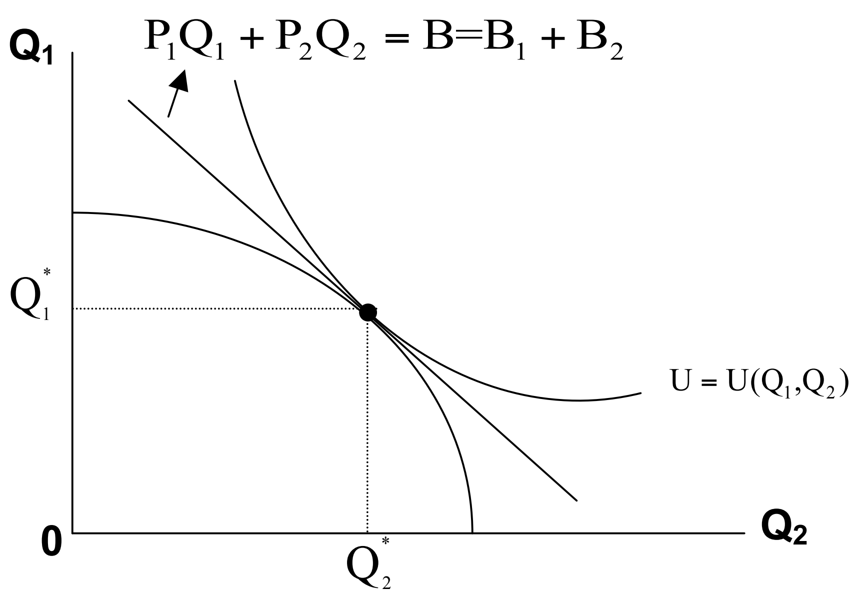


Figure (4)
The optimum solution for all three groups is the same, Q^*_1 and Q^*_2.

Conclusion

Close operational and financial alliances between airports and airlines would surely guarantee a success and provide financial benefits to the communities they serve. According to the International Air Transport Association (IATA) the aviation industry suffered net losses of more than $ 35 billion in the last four years from 2001 to 2004. United Airlines, US Airways and ATA Airlines are operating under Chapter 11 bankruptcy protection, and other carriers appear to be hovering on the edge. Knowing the practical limitation, the authors propose a theoretical framework for an airline-airport integrated operation optimization, in which all three stakeholders, the airlines, airports, and their customers are incorporated.

In the U.S., a majority of commercial airports are publicly owned, giving rise to different organizational structures similar to those associated with other countries in the European Union. In an ideal special case, the model provides for a practical integration of airports and commercial airlines into one unified enterprise, in which the main output would be a two-tier (and complementary) output. This integration, in a competitive environment, may enhance airport and airline efficiency, and reduce cost, which could facilitate airports to free up some resources for other activities.

References

Air Transport Association of America. Retrieved online 12/1/04, http://www.airlines.org/news/d.aspx?nid=9194

Airport System Development. (1984). U.S. Congress, Office of Technology

Assessment, OTA-STI-231, Washington, D. C.

American Association of Airport Executives (EEEA). (2003). Airport Rates and Charges, 2002-2003, 2, 8.

Blanchard, Oliver. (2000). *Macroeconomics*. Prentice Hall, 2nd edition, 453.

Federal Aviation Administration (FAA). (2002). FAA Report to Congress National Plan of Integrated Airport Systems (NPIAS) (2001 - 2005). Washington, DC Retrieved online 03/1/05. http://www.faa.gov/arp/planning/npias/npias2001/npias2001.htm

FAA. (2004). *Commercial Air Carriers*. Federal Aviation Administration. Washington, DC. Retrieved online: http://www.faa.gov/data_statistics/aviation/aerospace_forecasts/2004-2015/media/CHAPT3-04.pdf

Button, K. (2002). "Toward Truly Open Skies." *Regulation.* Washington, D.C., 25, 3, 12-17.

Hamzaee, R. & B. Vasigh. (2002). Total Factor Productivity Index (TFP) for Airport efficiency." Handbook of Airline Economics, McGraw Hill Publishing Co., 2nd edition, 469-

476.

Hamzaee, R.. & Vasigh, B. (2001). "Airport Efficiency." *AVMARK Aviation Economist Journal*. 10-12.

Hamzaee, R. & B. Vasigh, (2000). A simple revenue-cost oerspective on U.S. Airport Operations. *Journal of Air Transport Management*, Center for Transportation, George Mason University, and Elsevier Science Ltd., Oxford, U.K.

Hamzaee, R. & B. Vasigh, (1997). "An Applied Model of Airline Revenue Management." *Journal of Travel Research*, University of Colorado, Boulder, Colorado, Spring, 35(4), 64-68.

Henderson, J. & R. Quandt, (1976). *Microeconomic Theory: A Mathematical Approach, McGraw-Hill*.

International Air Transport Association, (2004). Global Press Briefing, Geneva.

Lehman Brothers, (2003). Combining Private Equity, Economic Development and Transportation, Overview of Transportation Public-Private Partnership Project Financing.

Johnson, K., (2001). "Munich International Boosts Hub with Terminal Expansion Project." *Airport*. 18 (44), 5.

Lovens, Pierre-François. (2003). Ryanair under fire at Charleroi. Institut des Sciences du Travail.

Vasigh, B. & R. Hamzaee, (1998). Comparative Analysis of Economic Performances of U.S. Commercial Airports." *Journal of Air Transport Management*, Center for Transportation, George Mason University, 4(4), 209-216.

Wall Street Journal, September 19, 2005, A1

SHEEP FLOCKS HEREDITY MODEL ALGORITHM FOR SOLVING JOB SHOP SCHEDULING PROBLEMS

M.Chandrasekaran, National Institute of Technology, India.
P.Asokan, National Institute of Technology, India.
S.Kumanan, National Institute of Technology, India.
T.Balamurugan, National Institute of Technology, India.

ABSTRACT

Job shop Scheduling (JSP) is a NP- hard combinatorial problem. In the job shop scheduling, a finite set of jobs is processed on a finite set of machines under certain constraints, such that the maximum completion time of the jobs is minimized. In this paper, an evolutionary computational algorithm based on sheep flocks (SFHM) is proposed and applied for job shop scheduling problem with the objective of minimizing the makespan and tested with 130 benchmark problems. The results of proposed algorithm are compared with Tabu Search Shifting Bottleneck (TSSB) procedure, Artificial Immune System (AIS) and as well as best solution of Shifting Bottleneck procedure (SB-GLS1) of Balas and Vazacopoulos. The performance of SFHM algorithm in job shop scheduling is found competent.

Keywords

Job shop scheduling, Sheep flocks heredity model algorithm, Benchmark Problems

Introduction

Job shop scheduling (JSP) has a very wide and well-developed engineering background, and it can be described as follows. A set of *m* machines and a set of *n* jobs are given; each job consists of a set of operations that have to be processed in a specified Sequence, each operation has to be processed on a definite machine and has a processing time which is deterministically known. A schedule defines the time intervals in which the operations are processed, but it is feasible only if it complies with the following constraints: each machine can process only one operation at a time and the operation sequence is respected for every job. The objective is to find the optimal schedule, i.e. schedule the operation order and starting time on each machine such that the makespan is minimal.

Job Shop Scheduling Problem

Normally, the entire job-shop scheduling problem consists of two types of constraints: sequence constraint and resource constraint (Yang, 2003). The first type states that two operations of a job cannot be processed at the same time. The second type states that no more than one job can be handled on a machine at the same time. Job-shop scheduling can be viewed as an optimization problem, bounded by both sequence and resource constraints. For a job-shop scheduling problem, each job may consist of different number of operations, subjected to some precedence restrictions. Commonly the processing orders of each job by all machines and the processing time of each operation are known and fixed. Once started operations cannot be interrupted. Assume job i (i=1,2,…n) requires processing by machine k(k=1,2,…m) exactly once in its operation sequence (thus, each job has m operations).Let p_{ik} is the processing time of job i on machine k, X_{ik} is the starting time of job i on machine k, q_{ijk} is the indicator which takes on a value of 1 if operation j of job i requires machine k, and zero otherwise. Y_{ihk} is the variable which takes on a value of 1 if job i precedes job h on machine k, and zero otherwise.
The objective function for the given Job Shop Scheduling is

$$\text{Minimize } Z = \sum_{k=1}^{m} q_{imk} (X_{ik} + p_{ik}) \qquad (i = 1,\ldots.n)$$

Subject to

a) ***Sequence constraint***

$$\sum_{k=1}^{m} q_{ijk} (X_{ik} + p_{ik}) \leq \sum_{k=1}^{m} q_{i,j+1,k} X_{ik}, \qquad (i=1,\ldots..n; j=1,\ldots\ldots m-1)$$

ie., for a given job i, the $(j+1)^{st}$ operation may not start before the j^{th} operation is completed.

b) ***Resource constraint***

$$X_{hk} - X_{ik} \geq p_{ik} - (H + p_{ik})(1 - Y_{ihk}),$$
$$X_{ik} - X_{hk} \geq p_{hk} - (H + p_{hk}) Y_{ihk},$$
$$\text{where } (i=1,\ldots..n; h=1,\ldots..n; k=1,\ldots\ldots m)$$

where H is a very large positive integer, chosen so that only one of the above constraints binding either for Y_{ihk} =1 or for Y_{ihk} = 0.

Solution Methodologies

Bruker (1995) and Garey (1976) show that the job shop scheduling is an NP-hard combinatorial problem and is therefore unlikely to be solvable in polynomial time. Carlier and Pinson (1989) and Bruker (1994) managed to generate optimal schedules using a branch-and-bound approach, but only for small size problems. Therefore, the effort has been directed at the development of efficient local search-based heuristics to solve practical size problems by Blazewicz (1996), and Anderson (1997). Shifting bottleneck heuristic method (Adam et al, 1988) is one of the heuristic method for minimizing makespan in a job shop problem. Bianco, Dellolmo, Giordani, and Speranza (1999) have minimized the makespan in a multimode multiprocessor shop-scheduling problem.

In the case of an enumerative procedure, the effort, time, and mathematical sophistication for finding and understanding a suitable method may be judged not worth the possible return.

There may be no method, which is suitable for the complex constraints of the real situation. The methods may be inapplicable in practice, owing to inappropriate computing time, deceptive solution quality, or lack of robustness (Pirlot, 1996). As a consequence, optimization methods may be ignored and some rules known as heuristics may be implemented. These heuristics are problem specific and not consistent.

In recent years, much attention has been devoted to heuristics like Simulated Annealing (SA), Tabu Search (TS), Genetic Algorithm (GA) and Neural Network (NN). They are applicable, in particular, for solving combinatorial optimization problems (Ponnambalam et al, 2000). Among the meta-heuristic algorithms, the Genetic Algorithm (GA), inspired by the process of Darwinian evolution, has been recognized as a general search strategy and an optimization method, which is often useful for attacking combinational problems. During the last decade GAs have been widely applied in many engineering fields, especially in the field of production scheduling. A GA exhibits parallelism, contains certain redundancy and historical information of past solutions, and is suitable for implementation on massively parallel architecture.

In contrast to other local search methods such as Simulated Annealing & Tabu search, which are based on handling one feasible solution. GA has been used with increasing frequency to address scheduling problems. However, it is not easy to regulate a GA's convergence, so a GA often suffers from premature convergence (Leung et al,1997; Yang et al, 1998; Rudolph et al, 1994), and it is also difficult to choose suitable parameters for a GA (Grefenstette, 1988). In Integrated Generation and Transmission Maintenance Scheduling area, Sheep Flocks Heredity Model algorithm has produced better results (Koichi Nara et al, 1999). Solving Job shop scheduling using SFHM algorithm needs attention.

Sheep flock heredity model algorithm **(SFHM)**

Let us consider the several separated flocks of sheep in a field (Koichi Nara et al, 1999) as shown in Figure 1.

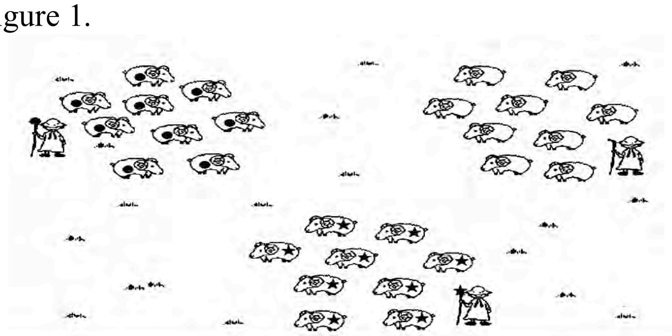


Figure 1. Flocks of sheep in the field

Normally, sheep in an each flock are living within their own flock under the control of shepherds. So, the genetic inheritance only occurs within the flock in other words, some special characteristics in one flock develop only within the flock by heredity, and the sheep with high fitness characteristics to their environment breed in the flock. In such a world, let us assume that two sheep flocks were occasionally mixed in a moment when shepherds looked aside as shown in Figure 2. Then, shepherd of the corresponding flocks run into the mixed flock, and separate the sheep as before. However, shepherds cannot distinguish their sheep originally they owned because the appearance of any sheep is the same. Therefore, several sheep of one flock are inevitably mixed with the other flocks as shown in Figure 3. Namely, the characteristics of the sheep in the neighboring flocks can be inherent to the sheep in other flocks in this occasion. Then, in the field, the flock of the sheep, which has better fitness characteristics to the field environment, breeds most.

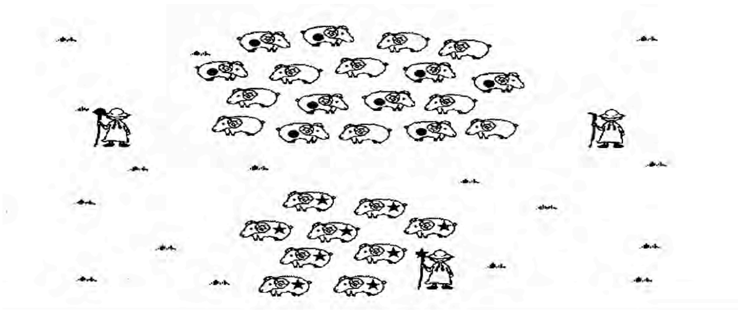


Figure 2. Mix of two flocks of sheep

Figure 3. New flocks of sheep in field

Natural Evolution to Multi- Stage Genetic Operation

Let us assume that a GA string can be divided into several sub-strings, and a length of each sub-string is the same (Nara, 1996). Then, we have the string structure as shown in Figure 4.

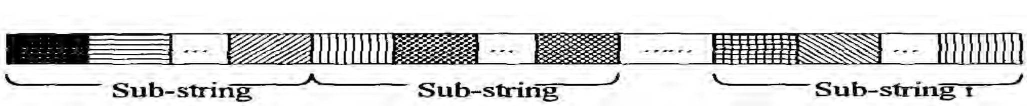


Figure 4(a) String structure when it is expanded.

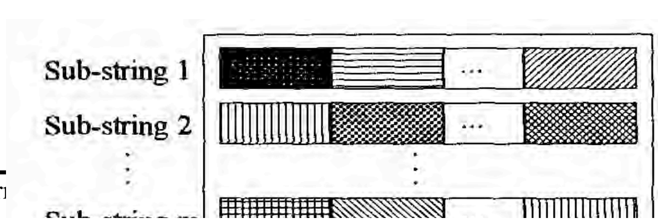


Figure 4(b). String when it is folded up in sub-string by sub-string.

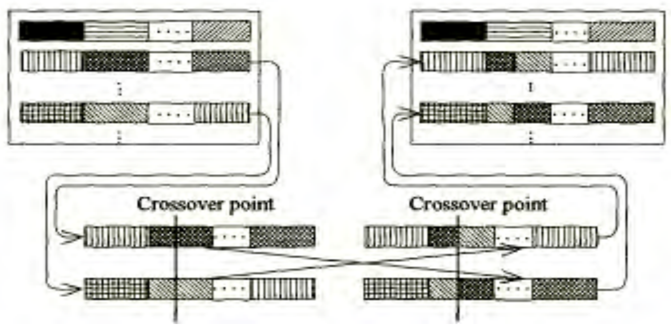


Figure 5. Genetic operation between strings

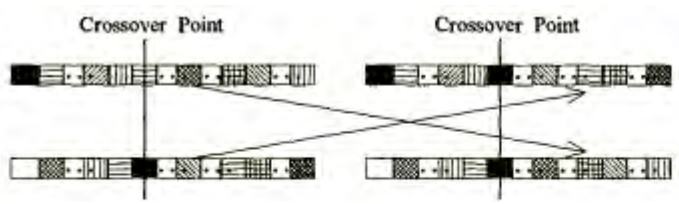


Figure 6. Genetic operation between sub-strings.

The natural evolution phenomenon of sheep flocks can be corresponded to the genetic operations of this type of string. For this kind of string, we can define the following two kinds of genetic operations. That is: (1) Normal genetic operations between strings as shown in Figure 5. (2) Genetic operations between sub-strings within one string as shown in Figure 6. This type of genetic operation to "multi-stage genetic operation". Correspondence of the elements of natural elements and proportionate multi stage genetic operation is shown in Table 1.

In sheep flocks heredity model algorithm special string structure, hierarchical genetic operations (crossover and mutation) are introduced. They are (1) sub-chromosome level genetic operation and (2) chromosome (global) level genetic operation. This hierarchical operation is referred to as "multi-stage genetic operation".

Natural evolution	Multi-stage genetic operation
Flock	String
Sheep	Sub-string
Mixed and separated	Chromosome level crossover
Inheritance with flock	Sub chromosome level crossover

Table 1. Correspondence of the elements

Application of SFHM Algorithm for Job shop scheduling problem

The flow chart of the proposed SFHM for job shop scheduling is shown in figure 7. The SFHM algorithm is implemented in C language on personal computer Pentium IV 2.4 GHz. The maximum number of iterations has been set to 100 X n, where n is the number of jobs. The algorithm is executed as below:

Begin
    Initialize the population

Stage 1:
        Select the parent
        Sub chromosome level crossover
        Set sub chromosome level crossover probability
        If population probability is less than or equal to sub chromosome level probability
                Perform sub chromosome level crossover
        Else retain the old sequences
        Sub chromosome level mutation
        Set sub chromosome mutation probability
        If population probability is less than or equal to sub chromosome mutation probability
                Perform sub chromosome level mutation
        Else retain the same sequences

Stage 2:
        Select two sequences from population
        Chromosome level crossover
        Set crossover probability
        If population probability is less than or equal to crossover probability
                Perform chromosome level crossover
        Else retain the same sequences
        Chromosome level mutation
        Set mutation probability
        If population probability is less than or equal to mutation probability
                Perform chromosome level mutation
        Else retain the same sequences
        End if terminal condition satisfied

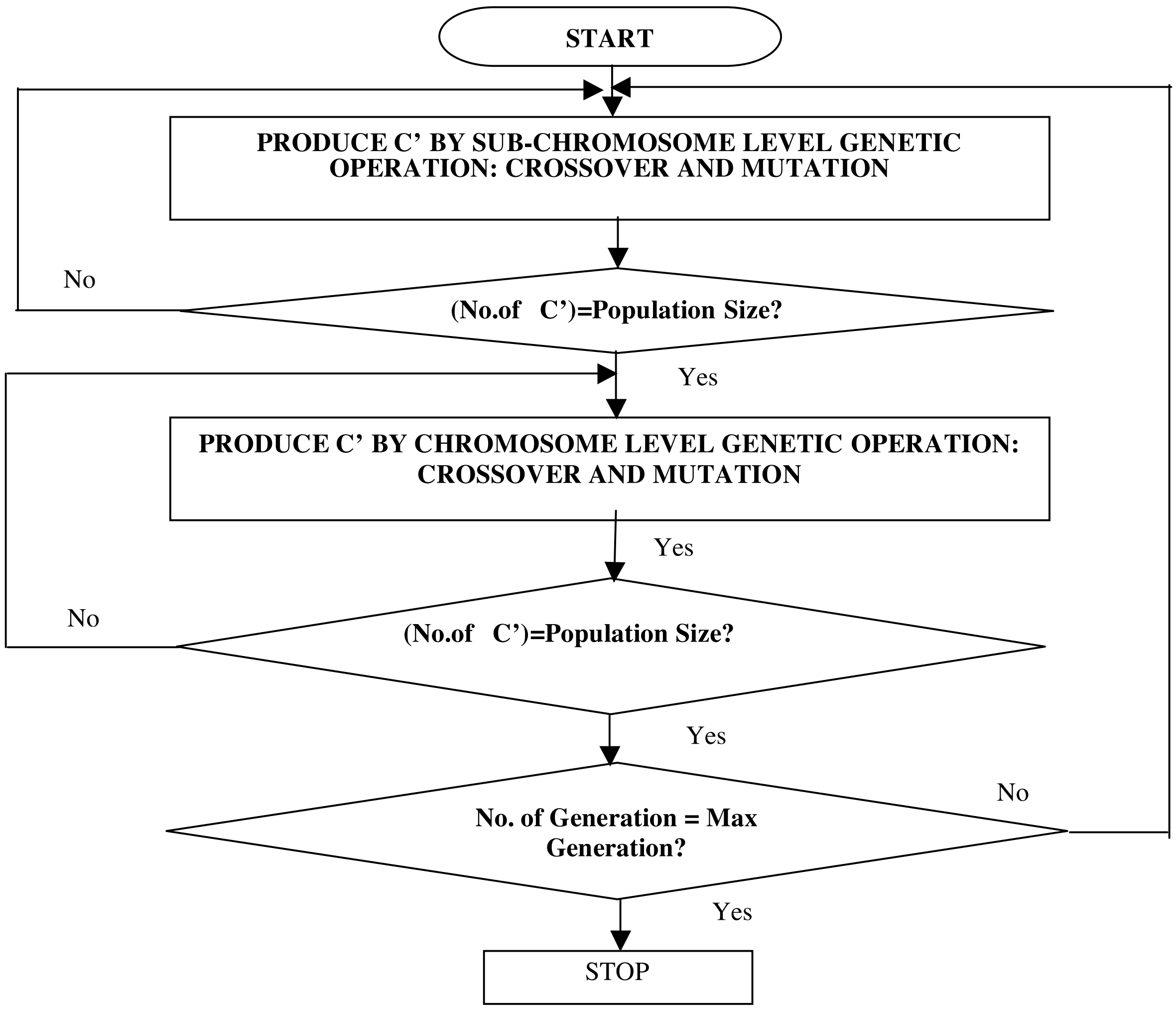


Figure 7. Flow chart of Sheep flocks heredity model algorithm for JSP

Case Example

Lawrence (LA16) Problem

It is 10-machines X 10 jobs problem. Table 2 and Table 3 shows the operation sequence and its corresponding processing time. The initial sequence generated randomly and the corresponding makespan for the sequence is given Table 4. With a crossover probability a second and a third sub chromosomes (sub string) as shown in the Table 5 are chosen randomly and crossover is performed.

1	6	9	8	7	2	0	4	3	5
4	2	5	9	0	7	1	8	6	3
3	2	8	1	4	9	7	6	0	5
1	3	2	7	8	9	6	0	5	4
2	0	5	6	7	1	4	9	3	8
2	3	5	9	4	6	0	8	1	7
3	2	0	1	9	8	6	5	4	7
1	0	3	4	6	9	8	5	2	7
4	2	8	5	3	7	1	6	9	0
8	9	2	4	3	0	7	6	1	5

Table 2. Operation Sequence job shop under study

21	71	16	52	26	34	53	21	55	95
55	31	98	79	12	66	42	77	77	39
34	64	62	19	92	79	43	54	83	37
87	69	87	38	24	83	41	93	77	60
98	44	25	75	43	49	96	77	17	79
35	76	28	10	61	9	95	35	7	95
16	59	46	91	43	50	52	59	28	27
45	87	41	20	54	43	14	9	39	71
33	37	66	33	26	8	28	89	42	78
69	81	94	96	27	69	45	78	74	84

Table 3. Processing Time of the jobs under study

7	8	6	9	4	10	5	1	3	2	Makespan = 1157
7	9	2	4	1	6	5	2	3	8	
9	10	7	5	8	4	1	10	6	3	
9	1	5	7	3	10	8	2	6	4	
2	3	7	9	8	10	1	5	4	6	
1	4	3	7	9	5	8	10	6	2	
6	10	5	1	9	7	4	2	3	8	
9	2	5	3	1	6	8	10	7	4	
3	6	8	4	2	9	10	7	5	1	
7	8	6	9	4	10	5	1	3	2	

Table 4. Initial sequence generated by SFHM with makespan

7	8	6	9	4	10	5	1	3	2
7	9	2	**4**	**1**	**6**	**5**	**2**	**3**	**8**
9	10	7	**5**	**8**	**4**	**1**	**10**	**6**	**3**
9	1	5	7	3	10	8	2	6	4
2	3	7	9	8	10	1	5	4	6
1	4	3	7	9	5	8	10	6	2
6	10	5	1	9	7	4	2	3	8
9	2	5	3	1	6	8	10	7	4
3	6	8	4	2	9	10	7	5	1
7	8	6	9	4	10	5	1	3	2

Table 5. Process of sub chromosome level crossover

Process of Sub Chromosome Level Inverse Mutation: Probability for this chromosome is less than process mutation probability. The second and Third sub strings are selected to perform this process. Each sub string 4 and 9 positions are chosen randomly to perform inverse mutation is shown in Table 6.

7	8	6	9	4	10	5	1	3	2
7	9	2	**3**	**2**	**5**	**6**	**1**	**4**	8
9	10	7	**6**	**10**	**1**	**4**	**8**	**5**	3
9	1	5	7	3	10	8	2	6	4
2	3	7	9	8	10	1	5	4	6
1	4	3	7	9	5	8	10	6	2
6	10	5	1	9	7	4	2	3	8
9	2	5	3	1	6	8	10	7	4
3	6	8	4	2	9	10	7	5	1
7	8	6	9	4	10	5	1	3	2

Table 6. Process of sub chromosome level Inverse Mutation in SFHM

Global Level Crossover Process: Crossover probability this sting is less than the process crossover probability. Sequences 1 and 6 are selected for crossover. At the position of 6thsub string is chosen as the crossover position which is shown in Table 7.

7	8	6	9	4	10	5	1	3	2
7	9	2	5	6	10	1	4	8	3
9	10	7	4	3	2	5	6	1	8
9	1	5	7	3	10	8	2	6	4
2	3	7	9	8	10	1	5	4	6
1	4	3	7	9	5	8	10	6	2
6	**2**	**7**	**9**	**4**	**3**	**1**	**10**	**8**	**5**
8	**1**	**7**	**3**	**2**	**9**	**10**	**6**	**4**	**5**
4	**8**	**10**	**7**	**3**	**5**	**9**	**1**	**2**	**6**
8	**4**	**10**	**2**	**7**	**9**	**1**	**5**	**6**	**3**

Table 7. Global Level Crossover Process in SFHM

Global Level Inverse Mutation Process: Probability for this string is less than process mutation probability. The 2nd and 6th sub string positions are randomly selected to perform inverse mutation. Inverse mutation process is shown in Table 8.

7	8	6	9	4	10	5	1	3	2
2	**6**	**10**	**8**	**5**	**9**	**7**	**3**	**4**	**1**
6	**4**	**5**	**1**	**10**	**8**	**9**	**7**	**3**	**2**
4	**6**	**2**	**8**	**10**	**3**	**7**	**5**	**1**	**9**
8	**1**	**6**	**5**	**2**	**3**	**4**	**7**	**10**	**9**
3	**8**	**4**	**1**	**10**	**6**	**5**	**2**	**9**	**7**
6	2	7	9	4	3	1	10	8	5
8	1	7	3	2	9	10	6	4	5
4	8	10	7	3	5	9	1	2	6
8	4	10	2	7	9	1	5	6	3

Table 8. Global Level Inverse Mutation Process in SFHM

***Final result obtained using SFHM algorithm*:** The terminal condition for the SFHM algorithm is set to specific number of iterations. Say it as 100. Then the solution is forwarded for scheduling. The Job shop schedule for LA16 problem under study is shown in Table 9.

7	8	3	4	9	6	5	2	10	1	Makespan = 945
5	7	4	2	1	3	9	8	6	10	
7	9	1	2	4	10	8	6	3	5	
1	9	7	5	8	3	4	6	10	2	
8	4	3	6	1	2	9	7	10	5	
3	10	9	7	6	4	2	5	8	1	
7	5	9	8	1	6	2	10	4	3	
5	10	7	2	6	4	8	3	9	1	
4	6	1	3	7	5	10	9	8	2	
6	2	9	3	8	10	5	1	7	4	

Table 9. Final Job shop schedule obtained using SFHM algorithm

Results And Discussion

The proposed Sheep flocks heredity model algorithm (SFHM) has been tested for 130 problem instances of various sizes collected in the following classes:

Class (i): Five instances denoted as (ORB1-ORB5) due to Applegate and Cook (1991) and Five instances denoted as (ABZ5-ABZ9) due to Adams et al. (1988).

Class (ii): Forty instances of eight different sizes (n X m = 10 X 5; 15 X 5; 20 X 5; 10 X 10; 15X10; 20 X 10; 30 X 10; 15 X 15) denoted as (LA01-LA40) due to Lawrence(1984).

Class (iii): Eighty instances of eight different size (n X m = 15 X 15; 20 X 15; 20 X 20; 30X15; 30X 20; 50 X 15; 50 X 20; 100 X 20) denoted as (TA1-TA80) by Taillard (1993).

The Relative Error RE (%) was calculated for all problem instances, as a percentage by which the solution obtained is above the optimum value (Opt) if it is known or the best-known lower bound (LB) (Pezzella et al, 2000).

$$RE (\%) = 100 X (UB – LB)/LB.$$

In Table 10, the solutions for Class (i) problems obtained from SFHM is compared with AIS and TSSB procedure and the mean relative error of SFHM (0.2601%) is found to be lower than the previous obtained results of (0.305%) AIS (Chandrasekaran et al, 2005) and (1.378%) from TSSB procedure (Pezzella et al, 2000). SFHM gives optimum bound value for 7 out of 10 problems where as AIS gives optimum bound value for 5 out of 10 problems and TSSB produces optimum value for 3 out of 10 problems. The graphical representation in Figure 9 shows the comparison of class (i) problem results obtained from SFHM with AIS and TSSB procedure.

In Table 11, the solutions for Class (ii) problems obtained from SFHM is compared with AIS and TSSB procedure and the mean relative error is 0.0268% in SFHM which is lower than previously obtained value of 0.048% of AIS and 0.14% by TSSB procedure. SFHM gives optimum bound value for 37 out of 40 problems where AIS gives optimum bound value for 36 out of 40 problems and TSSB produces optimum value for 33 out of 40 problems. The graphical representation in Figure 10 shows the comparison of class (ii) problem results obtained from SFHM with AIS and TSSB procedure. The mean relative error and average computing time is shown in Table 12.

In Table 13, the result obtained in SFHM is compared with AIS, TSSB and SB-GLS1 procedure. The mean relative error for class (iii) problem instances obtained by SFHM (0.751%) is lower than previously obtained results of 1.865% from AIS, 2.56% from TSSB procedure and 3.68% from SB-GLS1 procedure (Pezzella et al,2000). SFHM gives optimum bound value for 57 out of 80 problems where as AIS gives optimum bound value for 55 out of 80 problems, TSSB produces optimum value for 32 out of 80 problems and SB-GLS1 generates optimum value of 20 out of 80 problems. The graphical representation in Figure11-12 shows the comparison of class (iii) problem results obtained from SFHM with AIS, TSSB procedure and SB-GLS1 procedure. The mean relative error and average computing time is shown in Table 14.

Conclusion

Sheep flocks heredity model algorithm is applied for job shop scheduling problems. Since the method simulates both an inheritance among individual sheep and inheritance among sheep flocks, the method is referred as multi stage genetic operation. This algorithm has been tested with 130 benchmark problems. The percentage of deviations from lower bound is calculated. The results of sheep flocks heredity model algorithm is compared with Artificial Immune System, Tabu Search Shifting Bottleneck procedure and SB-GLS1 procedure. The proposed SFHM algorithm gives comparable results for most of the problems. The SFHM algorithm is another useful tool that does about as well as the algorithms to which it is compared.

References

Adams J., Balas E., Zawack D. (1988). The shifting bottleneck procedure for job shop scheduling, Management Science 34, 391-401.

Applegate D., Cook W. (1991). A computational study of the job shop scheduling problem, ORSA Journal on Computing 3, 149-156.

Bianco L., Dellolmo P., Giordani S., Speranza M.G. (1999). Minimizing makespan in a multimode multiprocessor shop scheduling problem. Naval Research Logistics, 46, 895–911.

Blazewicz J., Domschke W., & Pesch E. (1996). The job shop scheduling problem: conventional and new solution techniques. European Journal of Operational Research, 93, 1–33.

Bruker P. (1995). Scheduling Algorithms 2nd Edn (Berlin: Springer-Verlag).

Bruker P., Jurish B., and Sievers B. (1994). A Branch and bound algorithm for job shop problem, Discrete Applied Mathematics, 49, 107-127.

Carlier J., Pison E. (1989). An algorithm for solving the job shop problem, Management Science 35, 164-176.

Chandrasekaran M., Asokan P., Kumanan S., Balamurugan T. (2005). Solving job shop scheduling problem using Artificial Immune System, International journal of Advanced Manufacturing Technology, (Accepted) ***In Online***.

Garey M., Johnson D., Sethy R. (1976). The complexity of flow shop and job shop scheduling, Mathematics of Operations Research 1, 117-129.

Grefenstette J.J.,(1988). Optimization of control parameters for genetic algorithms, IEEE Transactions on Systems, Man and Cybernetics, 16(1), 122–128.

Koichi Nara, Tomomi Takeyama, Hyunchul Kim. (1999). A New Evolutionary Algorithm based on Sheep Flocks Heredity Model and Its Application to thermal generator maintenance scheduling problem, IEEE Transactions, Vol- VI , 503-508.

Lawrence S. (1984). Supplement to Resource constrained project scheduling: an experimental investigation of heuristic scheduling techniques, Tech. Rep., GSIA, Carnegie Mellon University.

Leung Y., Gao Y., and Xu Z.B. (1997). Degree of population diversity a perspective on premature convergence in genetic algorithms and its markov-chain analysis, IEEE Transactions on Neural Networks, 8(5), 1165–1176.

Nara K., Kim H. (1996). Evolutionary Computing by Multi-stage Genetic Operation and its Application to Optimal Scheduling problem, ***Proc. of*** ALIFE V, Japan.

Pezzella F., Emanuela Merelli, (2000). A tabu search method guided by shifting bottleneck for the job shop scheduling problem, European Journal of Operational Research 120, 297-310.

Pirlot M. (1996). General local search methods, European Journal of Operational Research, 92, 493–511.

Ponnambalam S.G., Aravindan P., Rajesh S. V. (2000). A Tabu Search Algorithm for Job Shop Scheduling, International Journal of Advanced Manufacturing Technology, 16, 765-771.

Rudolph G. (1994). Convergence properties of canonical genetic algorithms, IEEE Transactions on Neural Networks, 5(1), 96–101.

Taillard E. (1993). Benchmarks for basic scheduling problems, European Journal of Operational Research 64, 278-285.

Yang R. and Donglas I. (1998). Simple genetic algorithm with local turning-efficient global optimizing technique, Journal of Optimization Theory and Applications, 98(2), 449–465.

Yang S., Dingwei Wang. (2001). A new adaptive neural network and heuristics hybrid approach for job shop scheduling, Computers and operations Research 28, 955-971.

Notations Used

SFHM	-	best solution of SFHM algorithm
AIS	-	best solution of AIS algorithm
SB-GLS1	-	best solution of SB-GLS1 procedure of Balas and Vazacopoulos
TSSB	-	best solution of TSSB
m	-	number of machines
MRE	-	mean relative error in percent for a set of problems.
n	-	number of jobs
Opt (LB UB)	-	the optimal value of known best lower and upper bound, from OR-Library

$RE_{SB\text{-}GLS1}$	-	percent relative error by SB-GLS1
RE_{TSSB}	-	percent relative error by TSSB
RE_{SFHM}	-	percent relative error by AIS
RE	-	relative error in percent

Appendix

Problem	n	m	Opt (LB UB)	**SFHM**	**RE_{SFHM}**	CPU Time sec.	AIS	RE_{AIS}	CPU Time sec.	TSSB	RE_{TSSB}	CPU Time sec.
ORB1	10	10	1059	**1059**	**0**	25	1062	0.22	96	1064	0.47	82
ORB2	10	10	888	**888**	**0**	28	891	0.34	93	890	0.23	75
ORB3	10	10	1005	**1005**	**0**	19	1005	0	28	1013	0.8	87
ORB4	10	10	1005	**1005**	**0**	23	1005	0	35	1013	0.8	75
ORB5	10	10	887	**887**	**0**	29	889	0.23	98	887	0	81
ABZ5	10	10	1234	**1234**	**0**	22	1234	0	32	1234	0	75

ABZ6	10	10	943	**943**	**0**	16	943	0	37	943	0	80
ABZ7	20	15	656	**660**	**0.609**	178	666	1.52	256	666	1.52	200
ABZ8	20	15	(645 669)	**652**	**1.085**	143	669	0	118	678	5.12	205
ABZ9	20	15	(661 679)	**667**	**0.907**	159	684	0.74	242	693	4.84	195
Mean Relative Error (MRE)					**0.2601**			0.305			1.378	

Table 10. Results for ten instances (ORB1-ORB5) and (ABZ5-ABZ9) of Class (i) problems

Problem	n	m	Opt (LB UB)	SFHM	$\mathbf{RE_{SFHM}}$	AIS	RE_{AIS}	TSSB	RE_{TSSB}
LA01	10	5	666	**666**	**0**	666	0	666	0
LA02	10	5	655	**655**	**0**	655	0	655	0
LA03	10	5	597	**597**	**0**	601	0.67	597	0
LA04	10	5	590	**590**	**0**	590	0	590	0
LA05	10	5	593	**593**	**0**	593	0	593	0
LA06	15	5	926	**926**	**0**	926	0	926	0
LA07	15	5	890	**890**	**0**	890	0	890	0
LA08	15	5	863	**863**	**0**	863	0	863	0
LA09	15	5	951	**951**	**0**	951	0	951	0
LA10	15	5	958	**958**	**0**	958	0	958	0
LA11	20	5	1222	**1222**	**0**	1222	0	1222	0
LA12	20	5	1039	**1039**	**0**	1039	0	1069	0

LA13	20	5	1150	**1150**	**0**	1150	0	1150	0
LA14	20	5	1292	**1292**	**0**	1292	0	1292	0
LA15	20	5	1207	**1207**	**0**	1207	0	1207	0
LA16	10	10	945	**945**	**0**	945	0	945	0
LA17	10	10	784	**784**	**0**	784	0	784	0
LA18	10	10	848	**848**	**0**	848	0	848	0
LA19	10	10	842	**842**	**0**	842	0	842	0
LA20	10	10	902	**902**	**0**	902	0	902	0
LA21	15	10	1046	**1046**	**0**	1046	0	1046	0
LA22	15	10	927	**927**	**0**	929	0.21	927	0
LA23	15	10	1032	**1032**	**0**	1032	0	1032	0
LA24	15	10	935	**935**	**0**	935	0	938	0.32
LA25	15	10	977	**977**	**0**	977	0	979	0.2
LA26	20	10	1218	**1218**	**0**	1218	0	1218	0
LA27	20	10	1235	**1235**	**0**	1235	0	1235	0
LA28	20	10	1216	**1216**	**0**	1216	0	1216	0
LA29	20	10	(1142 1153)	**1145**	**0.623**	1153	0	1168	2.28
LA30	20	10	1355	**1355**	**0**	1355	0	1355	0
LA31	30	10	1784	**1784**	**0**	1784	0	1784	0
LA32	30	10	1850	**1850**	**0**	1850	0	1850	0
LA33	30	10	1719	**1719**	**0**	1719	0	1719	0
LA34	30	10	1721	**1721**	**0**	1721	0	1721	0
LA35	30	10	1888	**1888**	**0**	1888	0	1888	0
LA36	15	15	1268	**1268**	**0**	1273	0.39	1268	0

LA37	15	15	1397	**1401**	**0.286**	1397	0	1411	1
LA38	15	15	1196	**1196**	**0**	1196	0	1201	0.42
LA39	15	15	1233	**1233**	**0**	1233	0	1240	0.57
LA40	15	15	1222	**1224**	**0.164**	1230	0.65	1233	0.9
Mean Relative Error (MRE)					**0.0268**		0.048		0.14

Table 11. Results for forty instances (LA01-LA40) of Class (ii) problems

Problem	n	m	SFHM		AIS		TSSB	
			MRE	Tav	MRE	Tav	MRE	Tav
LA 01-05	10	5	0	7	0.12	15	0	9.8
LA 06-10	15	5	0	9.2	0	12	0	-
LA 11-15	20	5	0	17.6	0	22	0	-
LA 16-20	10	10	0	21.3	0	26	0	61.5
LA 21-25	15	10	0	59	0.04	83	0.1	115
LA 26-30	20	10	0.125	95	0	120	0.46	105
LA 31-35	30	10	0	111.3	0	155	0	-
LA 36-40	15	15	0.09	120	0.21	128	0.58	141
			0.0268		0.048		0.14	

Table 12. Comparison of Mean Relative Error and Computing Time of SFHM with AIS and

TSSB

Problem	n	m	Opt (LB UB)	**SFHM**	**RE$_{SFHM}$**	AIS	RE$_{AIS}$	TSSB	RE$_{TSSB}$	SB-GLS1	RE$_{SB-GLS1}$
TA1	15	15	1231	**1231**	**0**	1231	0	1241	0.812	1244	1.056
TA2	15	15	1244	**1244**	**0**	1244	0	1244	0	1255	0.884
TA3	15	15	(1206 1218)	**1206**	**0**	1206	0	1222	1.327	1225	1.575
TA4	15	15	(1170 1175)	**1177**	**0.171**	1170	0	1175	0.427	1191	1.795
TA5	15	15	(1210 1228)	**1213**	**0**	1215	0.41	1229	1.57	1256	3.802
TA6	15	15	(1210 1239)	**1212**	**0**	1210	0	1245	2.893	1247	3.058
TA7	15	15	(1223 1228)	**1229**	**0.081**	1223	0	1228	0.409	1244	1.717
TA8	15	15	(1187 1217)	**1153**	**0**	1187	0	1220	2.78	1222	2.949
TA9	15	15	(1247 1274)	**1231**	**0**	1297	0.4	1291	3.528	1291	3.528
TA10	15	15	1241	**1245**	**0.322**	1241	0	1250	0.725	1266	2.015
TA11	20	15	(1321 1364)	**1321**	**0**	1357	2.73	1371	3.785	1402	6.132
TA12	20	15	(1321 1367)	**1325**	**0**	1367	3.48	1379	4.391	1416	7.192
TA13	20	15	(1271 1350)	**1271**	**0**	1369	7.71	1362	7.16	1377	8.34
TA14	20	15	1345	**1345**	**0**	1345	0	1345	0	1361	1.19
TA15	20	15	(1293 1342)	**1298**	**0.388**	1348	4.25	1360	5.182	1383	6.961
TA16	20	15	(1300 1368)	**1310**	**0.769**	1351	3.92	1370	5.385	1418	9.077
TA17	20	15	(1458 1464)	**1468**	**0.685**	1458	0	1481	1.578	1519	4.184
TA18	20	15	(1369 1396)	**1369**	**0**	1412	3.14	1426	4.164	1433	4.675
TA19	20	15	(1276 1341)	**1288**	**0.904**	1336	4.7	1351	5.878	1376	7.837
TA20	20	15	(1316 1353)	**1366**	**3.80**	1347	2.36	1366	3.799	1398	6.231
TA21	20	20	(1539 1647)	**1558**	**1.235**	1649	7.147	1659	7.797	1692	9.942
TA22	20	20	(1511 1601)	**1522**	**0.729**	1627	7.67	1623	7.412	1638	8.405
TA23	20	20	(1472 1558)	**1477**	**0.339**	1556	5.7	1573	6.861	1594	8.288
TA24	20	20	(1602 1651)	**1602**	**0**	1624	1.37	1659	3.558	1714	6.991
TA25	20	20	(1504 1598)	**1544**	**2.65**	1580	5.05	1606	6.782	1631	8.444
TA26	20	20	(1539 1655)	**1539**	**0**	1672	8.64	1666	8.252	1698	10.331
TA27	20	20	(1616 1689)	**1662**	**2.846**	1688	4.45	1697	5.012	1722	6.559
TA28	20	20	(1591 1615)	**1617**	**1.63**	1602	0.69	1622	1.948	1653	3.897
TA29	20	20	(1514 1625)	**1556**	**3.434**	1583	4.55	1635	7.992	1639	8.256
TA30	20	20	(1473 1596)	**1496**	**1.561**	1573	6.78	1614	9.572	1621	10.048
TA31	30	15	(1764 1766)	**1767**	**0.169**	1764	0	1771	0.397	1809	2.551
TA32	30	15	(1774 1803)	**1796**	**1.873**	1824	2.81	1840	3.72	1840	3.72

TA33	30	15	(1778 1796)	**1778**	**0**	1829	2.87	1833	3.093	1844	3.712
TA34	30	15	(1828 1832)	**1833**	**0.274**	1841	0.71	1846	0.985	1898	3.829
TA35	30	15	2007	**2007**	**0**	2009	0.09	2007	0	2010	0.149
TA36	30	15	(1819 1823)	**1826**	**0.165**	1825	0.32	1825	0.33	1874	3.024
TA37	30	15	(1771 1784)	**1795**	**3.433**	1796	1.41	1813	2.372	1846	4.235
TA38	30	15	(1673 1681)	**1683**	**0.597**	1699	1.49	1697	1.435	1762	5.32
TA39	30	15	(1795 1798)	**1802**	**1.504**	1803	0.44	1815	1.114	1822	1.504
TA40	30	15	(1631 1686)	**1687**	**1.281**	1684	3.25	1725	5.763	1749	7.235
TA41	30	20	(1859 2023)	**1859**	**0**	2019	8.66	2045	10.005	2106	13.287
TA42	30	20	(1867 1961)	**1912**	**1.87**	1956	4.76	1979	5.999	2018	8.088
TA43	30	20	(1809 1879)	**1884**	**4.14**	1902	5.14	1898	4.92	1946	7.573
TA44	30	20	(1927 2003)	**1978**	**2.64**	1987	3.11	2036	5.656	2069	7.369
TA45	30	20	(1997 2005)	**2009**	**0.601**	2011	0.7	2021	1.202	2049	2.604
TA46	30	20	(1940 2033)	**1992**	**2.680**	1997	2.94	2047	5.515	2115	9.021
TA47	30	20	(1789 1920)	**1834**	**2.51**	1906	6.54	1938	8.329	1973	10.285
TA48	30	20	(1912 1973)	**1962**	**2.61**	1982	3.66	1996	4.393	2080	8.787
TA49	30	20	(1915 1991)	**1989**	**3.86**	1993	4.07	2013	5.117	2046	6.841
TA50	30	20	(1807 1951)	**1916**	**5.49**	1975	9.29	1975	9.297	2009	11.179
TA51	50	15	2760	**2760**	**0**	2760	0	2760	0	2760	0
TA52	50	15	2756	**2756**	**0**	2756	0	2756	0	2756	0
TA53	50	15	2717	**2719**	**0.073**	2717	0	2717	0	2717	0
TA54	50	15	2839	**2839**	**0**	2839	0	2839	0	2839	0
TA55	50	15	2679	**2679**	**0**	2681	0.07	2684	0.187	2683	0.149
TA56	50	15	2781	**2781**	**0**	2781	0	2781	0	2781	0
TA57	50	15	2943	**2943**	**0**	2943	0	2943	0	2943	0
TA58	50	15	2885	**2885**	**0**	2885	0	2885	0	2885	0
TA59	50	15	2655	**2655**	**0**	2655	0	2655	0	2657	0.075
TA60	50	15	2723	**2723**	**0**	2723	0	2723	0	2723	0
TA61	50	20	2868	**2868**	**0**	2868	0	2868	0	2891	0.802
TA62	50	20	(2869 2895)	**2869**	**0**	2895	0.9	2942	2.544	2962	3.242
TA63	50	20	2755	**2755**	**0**	2755	0	2755	0	2796	1.488
TA64	50	20	2702	**2702**	**0**	2702	0	2702	0	2726	0.888
TA65	50	20	2725	**2725**	**0**	2725	0	2725	0	2751	0.954
TA66	50	20	2845	**2845**	**0**	2845	0	2845	0	2845	0
TA67	50	20	(2825 2826)	**2830**	**0.176**	2842	0.6	2865	1.416	2841	0.566

TA68	50	20	2784	**2785**	**0.036**	2784	0	2784	0	2785	0.036
TA69	50	20	3071	**3071**	**0**	3075	0.13	3071	0	3071	0
TA70	50	20	2995	**2995**	**0**	2995	0	2995	0	3004	0.301
TA71	100	20	5464	**5464**	**0**	5464	0	5464	0	5464	0
TA72	100	20	5181	**5181**	**0**	5181	0	5181	0	5181	0
TA73	100	20	5568	**5568**	**0**	5568	0	5568	0	5568	0
TA74	100	20	5339	**5339**	**0**	5339	0	5339	0	5339	0
TA75	100	20	5392	**5392**	**0**	5396	0.07	5392	0	5392	0
TA76	100	20	5342	**5342**	**0**	5344	0.037	5342	0	5342	0
TA77	100	20	5436	**5436**	**0**	5436	0	5436	0	5436	0
TA78	100	20	5394	**5394**	**0**	5394	0	5394	0	5394	0
TA79	100	20	5358	**5358**	**0**	5358	0	5358	0	5358	0
TA80	100	20	5183	**5183**	**0**	5183	0	5183	0	5183	0
Total					**60.08**		149.214		204.8		294.4
Mean Relative Error (MRE)					**0.751**		1.865		2.56		3.68

Table 13. Results of Eighty instances of eight different size (n x m = 15 X 15; 20 X 15; 20 X 20; 30 X 15; 30 X 20; 50 X 15; 50 X 20; 100 X 20) of class (iii) problems.

Problem	n	m	SFHM		AIS		TSSB		SB-GLS1	
			MRE	Tav	MRE	Tav	MRE	Tav	MRE	Tav
TA 01-10	15	15	0.057	68	0.08	118	1.45	2175	2.24	57
TA 11-20	20	15	0.65	160	3.23	232	4.13	2526	6.18	113
TA 21-30	20	20	1.44	472	5.21	495	6.52	34910	8.12	165
TA 31-40	30	15	0.93	637	1.34	835	1.92	14133	3.53	175
TA 41-50	30	20	2.84	855	4.89	2331	6.04	11512	8.5	421
TA 51-60	50	15	0.07	379	0.01	665	0.02	421	0.02	152
TA 61-70	50	20	0.02	713	0.16	1315	0.39	6342	0.83	590
TA 71-80	100	20	0	254	0.01	1019	0	231	0	851
MRE			0.751		1.87		2.56		3.68	

Table 14. Comparison of Mean Relative Error and Computing Time of SFHM with AIS, TSSB and SB-GLS1 of Balas and Vazacopoulos

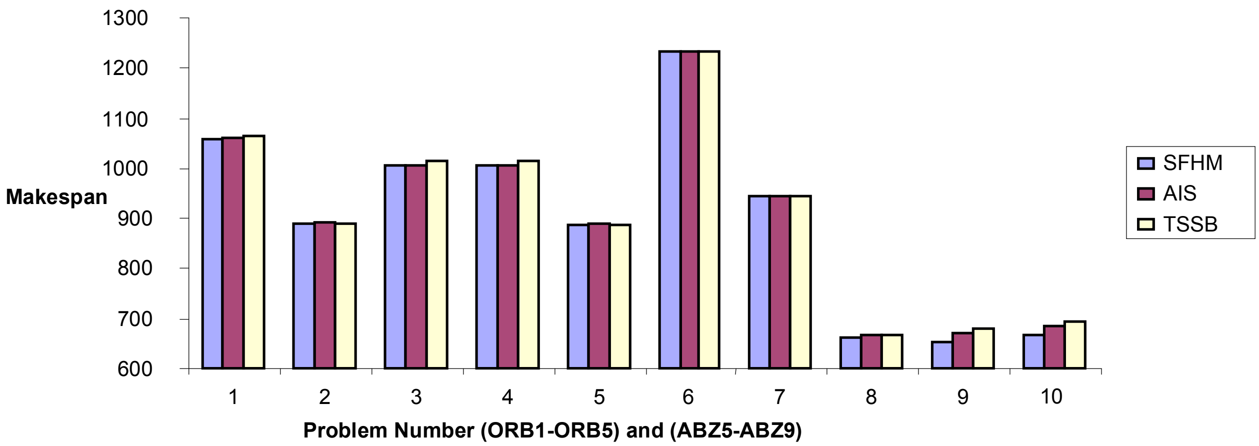


Figure.9. Graph shows that the results in SFHM algorithm compared with AIS, TSSB, and SB-GLS1 Procedure for problems (ORB1-ORB5) and (ABZ5 - ABZ9)

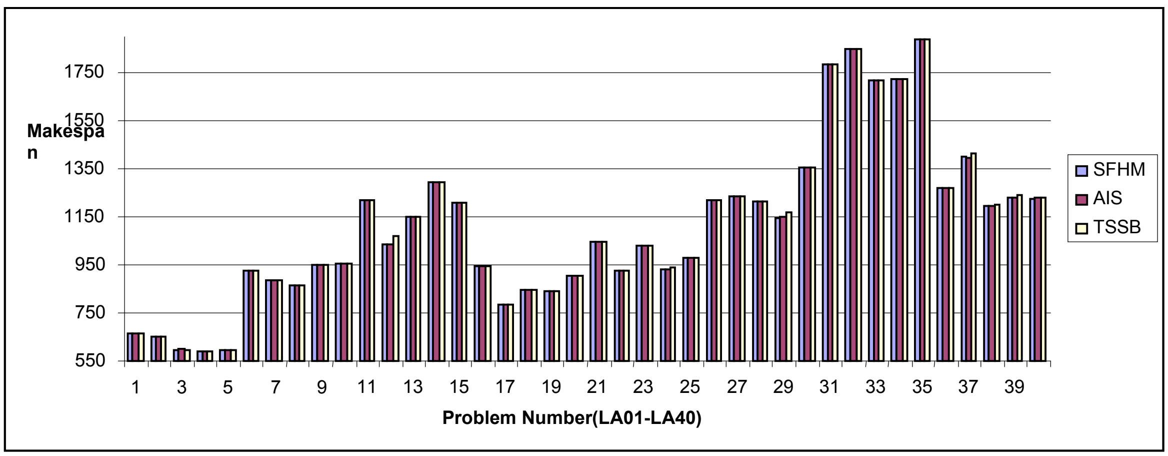


Figure.10. Graph shows that the results in SFHM algorithm compared with AIS and TSSB for problems LA-1 to LA-40

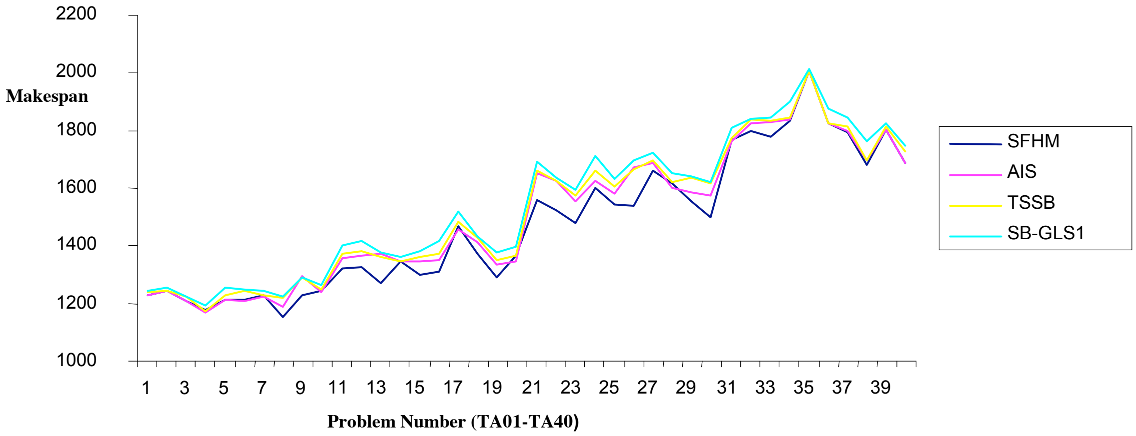


Figure.11. Graph shows that the results in SFHM algorithm compared with AIS, TSSB, and SB-GLS1 Procedure for problems TA-1 to TA-40

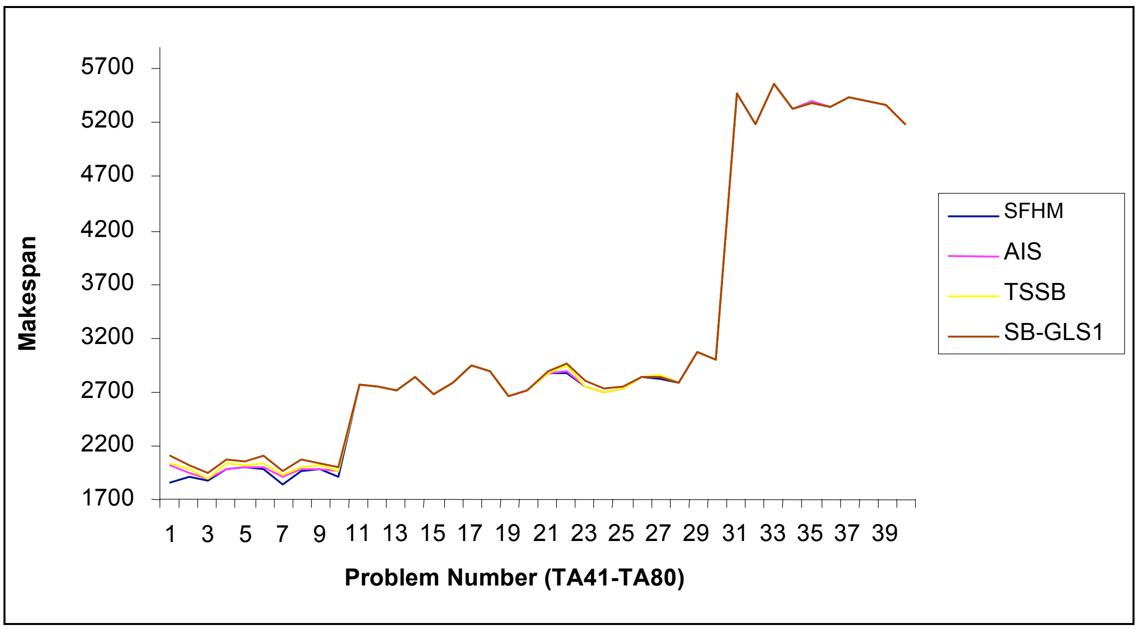


Figure 12. Graph shows that the results in SFHM algorithm compared with AIS, TSSB, and SB-GLS1 Procedure for problems TA-41 to TA-80

Contact Optimization: Asset Class Determination and Profile Risk

Michael P. Haydock, Walden University, USA

ABSTRACT

Social systems theory is concerned with the relationship, structure, and interdependence of groups. Inherent in the study of systems theory is the ability to segment a whole population into differentiated clusters where the characteristics and/or statistical properties will be minimized within the group, and, the differences in characteristics and/or statistical properties would be maximized between the groups.

This paper will utilize social systems theory to extend the segmentation process from just the clustering task into the five concepts of Von Bertalanffy. These concepts include arithmetic (binning); spatial (clustering); kinematic (topology); dynamic (state change); and biotic (valuation).

These processes are necessary to group customers into asset classes so that their behavior can be understood and forecasted. The portfolio of customer asset classes is also a way to value the worth of the firm. The concept is similar to the valuation and worth of a financial portfolio of investment instruments where the value of the portfolio is the discounted value of future cash flows. Decisions made about segments and their future behavior can lessen the investment risk inherent in promoting risky groups and is a core component of the contact economics concept.

Keywords

Social systems theory, Von Bertalanffy, decision sciences, asset class management

Asset Class Determination and Profile Risk

Systems theory is concerned with the problems of relationships, structures, and the interdependence of groups. Inherent in the study of systems theory is the ability to recognize the formation of groups and their relationships to the unified whole. Ludwig von Bertalanffy, the founder of modern systems theory, thought that the organized study of group behavior would have a broad application across many disciplines (Strauss, 2002). The target discipline and focused application of this paper is developing an understanding of how the marketing manager, operating in a direct marketing environment, can benefit from utilizing systems theory in the segmentation, profiling, and allocation of promotional dollars to their customer base.

Markowitz (1959) reflected on stocks, bonds, and money market instruments in his work on financial portfolio selection. Each of these instruments has characteristics such as durations, yields, risks, and return properties that are very different than the others. The wisdom for diversification within a financial portfolio has applicability to the task of diversifying customer investments by properly constructing customer portfolios. The marketing executive in like fashion groups customers into relatively homogeneous segments (or asset classes) to be studied. As Markowitz described, you need a little of each asset class in the portfolio if maximizing returns and minimizing acceptable risk is the investors objective. The characteristics of customer asset classes would differ on dimensions such as spend, treatment productivity, length of time as a customer, products bought, and a risk adjusted financial return profile, as an example. The dispositions of the asset classes are different, and they subsequently require different marketing treatments.

Von Bertalanffy was a biologist by training and approached the human social problem much in the same way that Markowitz approached the portfolio challenge, each decomposed the parts into segments or clusters and dealt with the individuality of the segments, but always with an eye on the whole. Using statistical or data mining methods, customer cluster types are segmented and isolated. How successfully these clusters are developed and profiled will determine how successful the marketing executive will be at profitably allocating promotional investment dollars and affecting response for the firms good and services. The risk in not segmenting properly is the missed opportunity to diversify and categorize the firm's most important asset, their customers.

The development and determination of which customers belong to which asset classes therefore is one of the more crucial tasks and ultimately determines how the downstream contact economics system treats any individual customer. The link between social systems theory and the tasks of segmentation, pattern detection, and classification possesses the opportunity to re-orient how data mining tools are used on human clusters. This paper will attempt to answer the marketing manager's questions regarding customer clusters, such as are they considered an open or closed system? What does one learn about the social system from the action of segmentation? Are these social systems self organizing in nature?

If the role of segmentation is to draw the boundaries between clusters, how is knowledge and insight provided in the process of understanding this social system? What does the marketing manager do with the information provided by segmentation once the customer clusters are formed? What attributes should the marketing manager use to describe the social system? What computational tools are at the marketing manager's disposal to assist in the task of clustering in multi-dimensional space?

In order to make improvements in the understanding and treatment of these consumer based social groups the task of customer segmentation should be viewed as part of an overall plan of improvement in the contact economics process. This process, illustrated in Figure 1, is referred to as the contact optimization management system and was designed to reduce the risk in managing long term customer relationships and to describe the economics behind promotional treatment decisions. A key component in the process is the customer segmentation function which classifies individual customers into asset classes where financial resources are allocated based on their risk and return profiles. The components of the risk management model covered in this work are highlighted with a red oval in Figure 1.

Any attempt to improve on an industry problem of this magnitude should begin with a good understanding of consumer segmentation processes relative to the theories of social systems and how those theories translate into today's consumer segmentation models. The premise is that consumer segmentation modeling through the use of data mining methods is a complex and somewhat formidable subject that can benefit greatly from a solid grounding in social systems theory. This paper will leverage experiences in the retail apparel direct marketing industry as examples throughout to illustrate the concepts and hopefully provide some insight into some of the complexity in modeling social groups.

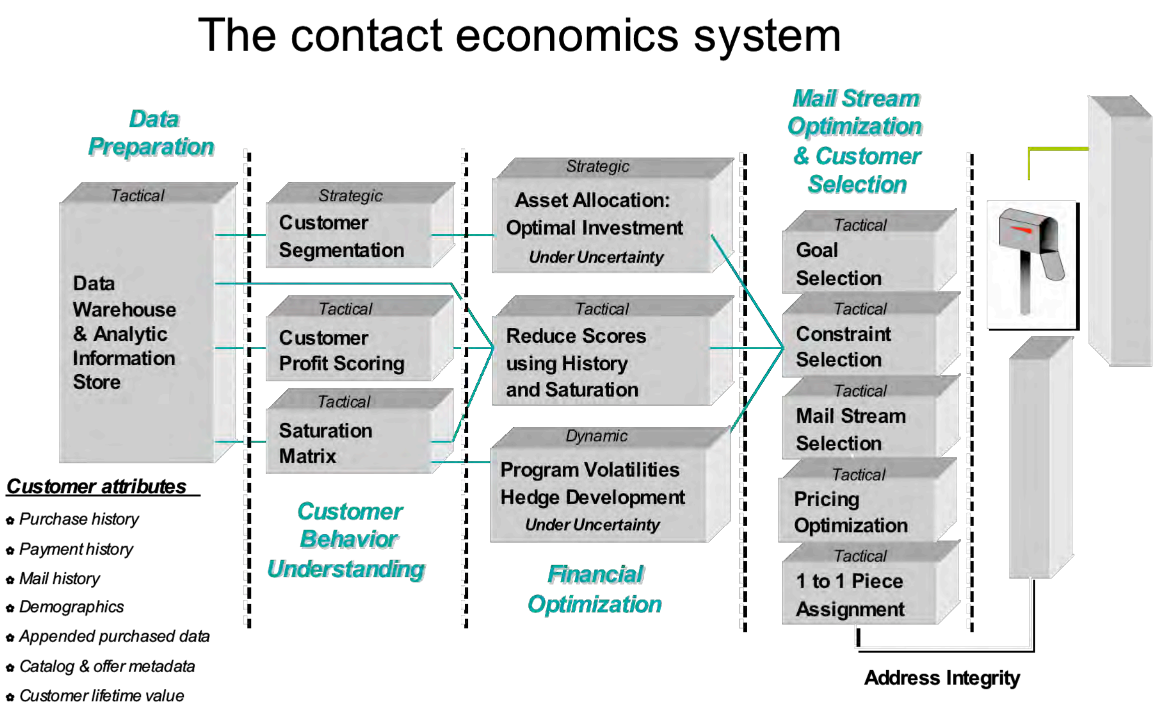


Figure 1. The contact optimization risk management system

From Social Systems to Asset Classes

Social systems are the most complex systems to define accurately. The buying behavior of individuals within these social systems may be among the most difficult to understand and model from a marketing context. The pursuit of accurate behavior modeling is something Sheth (1967) likens to the search for "inner peace", very difficult to define, articulate, or explain. Portfolio analysts in finance find the task of predicting the individual behavior of financial instruments equally as daunting, and have resorted to developing general "classes" of like instruments with similar characteristics as a proxy to understanding the market related to that class as a whole (Sharpe & Alexander, 1990).

How one chooses to allocate a set of investments between different types of assets (that is, between "asset classes") is the most important decision the investor can make when it comes to determining whether or not over time the firm will earn the minimum rate of return you need to meet your goals. Unfortunately, this "asset allocation" decision is one that most investors don't spend nearly enough time thinking about before they make it.

An asset class is therefore a group of securities that have more in common with the characteristics of members within the group than they do with members outside the group. Another way that could be put is to say that across their respective attribute space, the best asset class segmentation will have a minimum of differences within a group, and a maximum of difference between the groups. In the context of social systems theory, the definition of a statistically good cluster has also just been defined.

Sharpe & Alexander (1990) go on to state that because the riskiness of an asset class depends not only on how variable its return attributes are relative to their historical average (that is, their standard deviation), but risk also depends on the extent and direction in which the asset class's returns vary with those of other asset classes (that is, their correlations). For example, an asset class with a relatively low rate of return still might be a very good one to hold in a portfolio if those returns tended to go up when other asset classes' returns went down. Again, in the context of customer contact optimization, how a group is profiled relative to investment risk depends a lot on their returns relative to their average, and how they correlate to the investment returns of other groups.

So it seems that clustering assets is not enough, attribution and measurement must be part of the analysis. In fact the investor in financial securities is interested in the past as a precursor to the future, but is also keenly aware of how important forecasting the direction of expected returns of the group means to overall (the whole) portfolio performance. The investor would also care a lot about the transitions (inflows and outflows) experienced by the asset classes as well as the probabilities of that transition occurring and what that may mean for the investment potential of that asset class looking forward. So, the commercially available clustering capabilities available to the marketing investor provide for a segmentation in *n*-dimensional attribute space, but leave the determination of the returns forecast, risk profile, and investment strategy for the marketing manager to somehow adequately resolve.

From Asset Classes to Social Systems

Shepard (2003, p. 39) states that the fundamental goal of direct marketers with respect to segmentation is to "identify groups, segments, or clusters from a marketing perspective that are meaningfully different from each other". But is this really the marketer's *only* goal in the segmentation process? Is the marketer's role just that of developing efficient clusters? Isn't the marketer charged as well with understanding the social systems that make up the clusters? To do that we have to return to what Von Bertalanffy thought about the importance of understanding the whole and its parts (Strauss, 2002).

Von Bertalanffy described a social system as unity in the form of principles and laws concerning what made up the organization. Five conceptual clusters are described (Strauss, 2002):

- An arithmetic cluster, defining the one and the many,
- A spatial cluster, defining the boundaries of the groups within the whole,
- A kinematic cluster, defining the transitions and movement of the groups within the whole,
- A dynamic cluster, which defines the relationship between the structure of a system and its behavior looking forward, and
- A biotic cluster, which defines the health of the system looking forward, in business terms, this would equate to its economic health.

The connection between these clusters is an important example of the whole (the social system under investigation) being understood as a result of the study of its parts. This unity of concept cluster components within a multiplicity of the whole is a core learning that may differentiate the true ability to understand the nature of a social group through investigation of data. Could social systems theory provide the marketer with a more powerful way to *know* the segments of the market they are charged with? Figure 2 outlines the conceptual clusters that Von Bertalanffy describes as a set of processes. The vertical demarcation line (in red) separates the differences in how the processes are applied today by the typical marketer (Shepard, 2003) and what additional processes are implicit in using social systems theory in order to *know* about a population.

The first three process blocks illustrated in Figure 2 represent from a process point of view the current state of the art in segmentation and clustering methods that are captured in the literature on data mining. An excellent treatise on the topic of data mining and knowledge discovery from large scale databases (KDD) by Fayyad & Stolorz (1997) primarily discuss the use of techniques which can interrogate large scale data bases without a specific query in mind, but with the target of understanding the structure or hidden patterns within the data. This computer driven exploration approach to the data is very different from human-driven exploration in that the computer is not forming a hypothesis about the data as the human analyst must to execute a directed query.

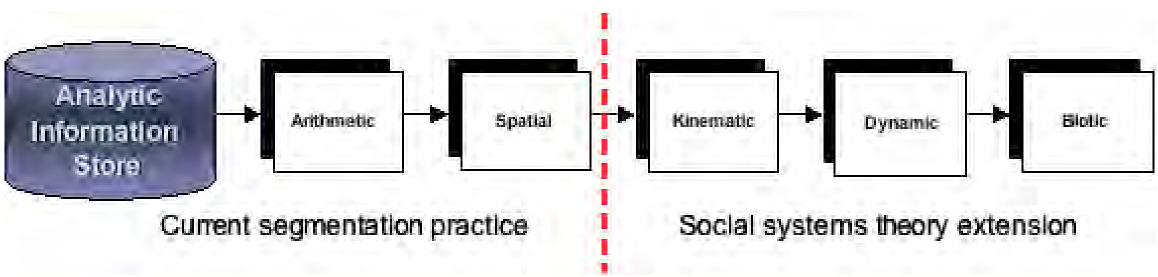


Figure 2. Current segmentation practices and social systems theory extensions.

Data mining is therefore one way within the overall set of social system processes of extracting useful information from data. The use of the term data mining often conjures up the “blind” exploration of data in order to gather information, but normal data mining processes are quite iterative (Fayyad & Stolorz, 1997). The role of data mining in pattern detection and classification is to accumulate the collection of observations that are connected in space or time, or both, and to discern the structure of an underlying pattern (Schurmann, 1996).

These patterns should exhibit certain regularities in such a way that a concept can be developed about the data. In Fayyad & Stolorz’ article the emphasis is placed on describing the modeling aspects that primarily lead to clustering and segmentation. Hosking, Pednault & Sudan (1997) describe a from statistical perspective the various processes that data mining can utilize to discern patterns within the data using classification as a very good example. Model selection is emphasized and initial treatment of the data is discussed.

This notion of model selection is articulated by Michaud (1997) who eloquently describes various clustering algorithms which create partitioning cuts within the data as a way to aggregate data into discernable groupings. Michaud proposes the Condorcet algorithm for voting choice as an improvement to current clustering technologies. Zait & Messatfa (1997) compares and contrasts various clustering and segmentation methods giving a very well thought out method for comparing the efficiency of various data mining algorithms in detecting patterns in large scale databases. But in no sense did these accomplished data mining experts articulate what to do with the clustering and segmentation results once completed.

But, did the clustering results clearly provide information useful to the experienced marketer? A marketing manager utilizing these tools against a massive amount of customer records could indeed achieve a sense of the separable and differentiated segments of the firm's marketplace. The use of the data mining tools though pre-supposes that the marketing manager could discern other information that was valuable about the groups looking forward.

Should the marketing manager also care about the kinematics of the newly formed clusters? The early Greek philosopher Heraclitus stated that one can not step into the same river twice, as fresh water would have since flowed over it. His thoughts were directed toward the tension between stability and change. Kinematic behavior about a group may give the marketing manager a most relevant insight into the areas of stability and change about the social system under investigation. Marketers want movement between the groups as a result of administering treatment, but there is a desire for stabilizing features of groups as well.

The flow between groups could be viewed as transitions over the topology of a network of segments, all connected making up the whole. Is there stability at the base of whatever is in transition? This is Von Bertalanffy's concept of the unbreakable foundational coherence between constancy and change. Without this extended knowledge of the kinetic topology of the group, how would the marketing manager know which marketing treatments to apply to the social system?

How should the marketing manager look forward in time to determine how these segments will organize themselves in future states? The trajectory of the social system may no longer follow the determinate path of prior states and may enter the domain of uncertainty. The marketing manager should extract from the segmentation process the pattern of change taking place over multi-periods and be cognizant of the force and intensity of the dynamics taking place.

Dynamic modeling of the prior transition states with Bayesian posterior outputs at each time step could help the marketing decision maker understand how actions in one time period tend to affect the future state of the social system. Bayesian methods provide the probabilities of conditional state transitions at each future time step (Harsanyi, 1967).

Traditional segmentation techniques make no reference to forecasting trajectories, but only in summarizing the current state of the past. One potential view of this process is that the marketing manager is playing a "game" against a group, or a market. The manager is charged with the first move that usually comes in the form of a marketing treatment and must decide on the intensity of the treatment. A set of payoffs occur based on the adverse state of the system and the nature of the investment risk taken (treatment intensity). The output from this dynamic process can illustrate the direction of growth the firm's customer file takes in the future and is immensely valuable in the allocation of financial resources to these asset classes (Haydock & Bibelnieks, 1999).

However unstable the anticipated dynamic state change process is, the biotic state of the cluster should provide an irreducible balance by valuing the social groups in each transitional state in terms of the opportunity that exists for each group at each state. These financial opportunities are again balanced by the optimal investment at that opportunity (Haydock & Bibelnieks, 1999).

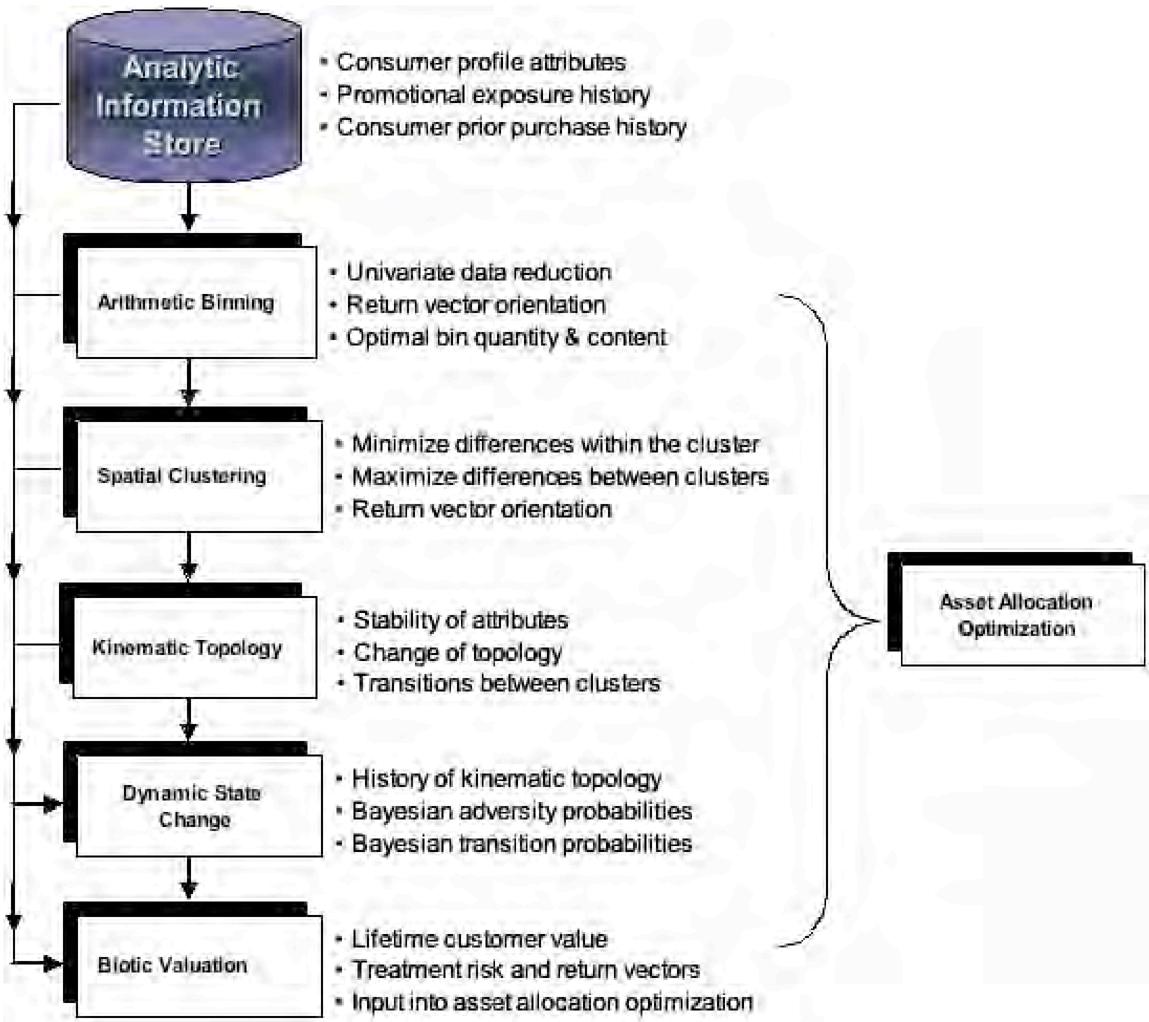


Figure 3. Social systems segmentation process and optimal allocation of assets.

Asset class investment optimization is a process that balances this tension between dynamic change and biotic stability and is a function of trading off the characteristics contained in the asset classes, and adding the attributes of the kinematic topology, the dynamic state change, and the biotic valuation of the groups. The proposed process for developing a new segmentation scheme would include the suite of capabilities as illustrated in Figure 3.

Strauss (2002) correctly identifies that Von Bertalanffy realized that groups consisting of biotic entities that exist in a thermodynamic sense, are to be considered open systems. Strauss continues with this thought and states that the relevant systems to watch are those that exist in a state of dynamic non-equilibrium with an enduring energy flow. In relating this thought to the direct marketing manager's challenge of optimally treating a customer base with promotional resources (energy flow), the segments to watch are those in the most state of attribute and financial flux. The proposed process illustrated in Figure 3 will certainly take into account the traditional segmentation outputs of number and space, but will enhance the customer knowledge effort with the addition of movement, forecast, and valuation.

The next five sections of this will decompose each of the five parts of the proposed social systems segmentation process and articulate how that component should function, compare the proposed methods with traditional methods used by direct marketing managers, and show how all the concept parts work as a systems whole.

Arithmetic Binning: Unsupervised Univariate Clustering

For any given continuous univariate variable with a set of observations in the form: $v_1, v_2, v_3, ... v_n$, a statistically optimal cluster can be developed that can serve as a proxy for the detailed data where the elements of the cluster are statistically minimized and separable from the other univariate clusters. This type of cluster will be referred to as a bin, and the process of creating the bin will be referred to as arithmetic binning.

The process of selecting these bins may be either supervised or unsupervised. In a supervised binning activity the number of bins and the bin boundaries, which may be thought of as the left and right side of a histogram, are determined a priori by the analyst (Zhang & Cheng, 2003). In an unsupervised binning activity the number of bins and the boundaries are self-selected based on the statistics of the univariate variable being explored. In researching the literature (Roederer, Treister, Moore, & Herzenberg, 2001) and (Zhang & Cheng, 2003) it appears that supervised learning is the norm in binning activities and related commercial software.

The advantage of supervised learning is primarily ease of use and direct control over the content. There may be applications where the analyst with domain expertise prefers that there be a certain quantity of observations contained in the bin (Roederer et al., 2001). Other applications may set an upper bound on the number of bins in order that they have certain statistical properties (Zhang & Cheng, 2003).

The disadvantages of supervised binning would be:

1. There may be more statistically optimal bin configurations than any heuristic applied by the analyst.
2. The configuration and bin quantity are not data driven, but is analyst determined.
3. The classification of bin observations derived from a training set are rule based, versus data based where in large data sets could lead to overlapping bin distributions (Roederer et al., 2001).

The proposed binning application would be referred to as unsupervised. This procedure is directly applicable to use in the direct marketing environment where customer attributes are quite often continuous variables. Examples of these descriptive attributes would include the number of purchases, time as customer, lifetime amount spent, and specific merchandise purchased.

Most direct marketers accumulate extensive amounts of customer data that is eventually placed in an enterprise data warehouse (Samli, Pohlen, & Bozovic, 2002). These data are collected in a myriad number of ways such as point-of-purchase transactions at store locations, e-commerce transactions that occur over the Internet, and call center transactions while shopping a catalog. These transactions would constitute the raw behavioral observations. Before this data can be binned a series of transformations should occur in a pre-processing step to provide a measurement vector for the univariate attribute under investigation (Schurmann, 1996).

The process of taking raw data and converting the data into these transformations or measurement features are determined by the experience of the marketing analyst and their familiarity of the domain that they are applied to. The notion of measurements is used to describe this transformation process. The true target in developing these measurements is in the development of patterns and classification features. The treatment of the data into measurement features will not be handled in its entirety in this work but is clearly an area targeted for further research. The data and examples used in this work would have already been treated for optimal measurement features.

Figure 4 illustrates the measurement of statistical properties inside a bin using the concept of unsupervised optimal binning where *N* number of bins is created with values that tightly couple the observations in each bin. The first objective would be to minimize the Euclidean distance from the centroid or mean of each bin of observations that are classified into that bin. The Euclidean distance would be given by (Borowski & Borwein, 1989):

$$d(x,\mu) = \sqrt{\sum_{i=1}^{n}(x_i - \mu)^2} .$$

Where:

d = distance between two points in univariate space,
x = observation,
i= the ith observation,
μ = Mean of the binned series,
n = the number of observations.

A secondary objective would be to prevent any overlap of bin classification space between any two pairs of bins. A third objective is to separate bin classification into observations that have a positive response function and observations that have a null response function. The response function would be some enumeration of a return vector such as merchandise purchases. A fourth objective would be to insure the goodness of fit within the bins by computing Pearson's chi-square test to determine how well the data supports assumptions about the distribution of the univariate random variable under investigation (Aczel & Sounderpandian, 2002).

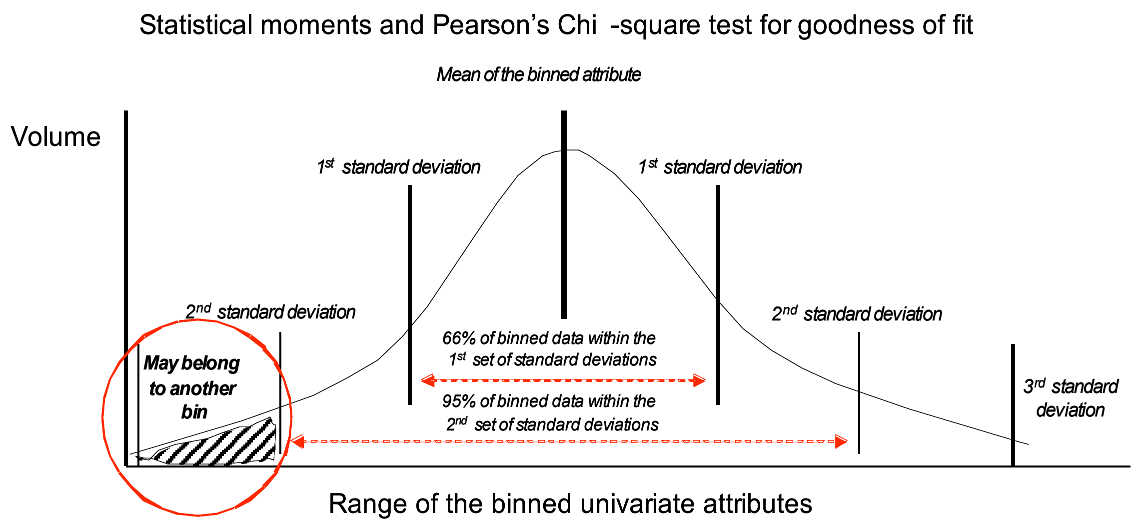


Figure 4. The attributes of an unsupervised binning classification process.

A fifth and ancillary objective would be to determine a method for the creation of a training set such that a certain number of randomly selected observations could create the bin topology and through classification techniques, have subsequent observations gated into the correct bin. This objective would be required where there exists a large amount of data and processing speeds become a concern during the arithmetic binning task.

An arithmetic bin could therefore be thought of as a pattern which has measurement (v) and an identity (w). So, each pattern = [v,w] where v is the collection of observations and w is the identity behind those observations (Schurmann, 1996). Whereas v may be drawn from millions of observations, w will be the finite set of bins optimizing the statistical features of v.

Schurmann (1996) describes that the detection of patterns which form an individual bin is therefore the process of inferring w from v. In a sense, v and w can be thought of as existing in two different spaces. One is in measurement space (v) and the other (w) being in identity space. Pattern classification with respect to binning is the process of mapping observation v to bin identity w when only the observation (v) is known. Figure 5 illustrates this mapping while Figure 4 accumulates what must be taken into account about what is known concerning the variability of v around the centroid of a given bin (w). This mapping is therefore a clear task that arithmetic binning must execute in order to understand the random processes surrounding pattern classification of v into w.

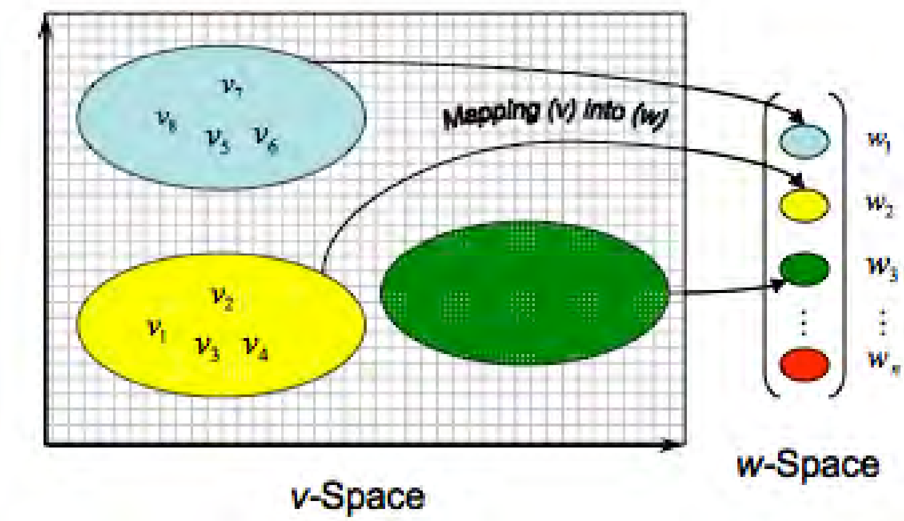


Figure 5. Pattern classification process mapping observations (v) into bins (w).

The process steps to create the arithmetic bins would be as follows:

- Extract and treat the raw observations by column vectors. This treatment should include separable measurement between responders and non-responders in the population.
- Determine the optimal number of bins through a pre-processing step.
- Bin the data with the objective of minimizing the Euclidian distances of the observations around the mean value of the bin.
- Classify each incoming observation by managing the mean, standard deviation, and covariance's in order to prevent overlap between bins.
- Utilize a Chi-square test for normality of the distribution within the bins.
- Adjust bin content until all bins are at statistical equilibrium.

Spatial Clustering: Unsupervised Multivariate Segmentation

Segmentation of a given population into similar groupings of attributes has many applications in the area of contact economics. The goal of segmentation from a social systems theory point of view is to gain knowledge into some attribute structure resident in the data of the population. Insight into this data is necessary to deploy the optimal treatment strategy to the members of the group. A principal of marketing is that differentiation between groups leads to optimal allocation of assets and maximizes profits of the firm (Kotler, 1994). The more the marketing manager knows about the group, or social system, the more customized the treatment can be.

Michaud (1997) states that in general it is not possible to define what it means to be "similar" in terms of the given attributes that defines a segment. Michaud also states that comparing one segment with another is also very difficult and that comparisons are usually application and domain specific. The difficulty lies in the use of multiple attributes to describe a customer and therefore the segment that customer is classified into.

Figure 6 illustrates the multi-dimensional aspect of the problem. The binning problem described previously had the advantage of clustering around one attribute. Contrasting Figure 5 with Figure 6 it is immediately apparent that the complexity of the problem has dramatically increased. The process of segmentation therefore could be thought of as binning the bins, but with very specific objectives in mind.

Such complexities inherent in multivariate segmentation have been explored aggressively over the past 10 years by commercial companies and data mining solutions do exist. There are three major objectives of the segmentation exercise:

1. To minimize the differences within the segments;
2. To maximizing the differences between the segments; and
3. To allow for a measure of financial return to enter into the segmentation.

Table 1 organizes some of the more widely used commercially available data mining software into types of data mining methods or capabilities and companies that market the software (Haughton et al., 2003). It would be very difficult to try to rank order capabilities from each company as to features since different companies excel at different features and some features are especially preferred by specific industries. Each of the companies represented in Table 1 have at least some capability in each of the method areas. The Open Source would not necessarily be considered a commercial entity as of yet, but could become an important new source of quality data mining codes.

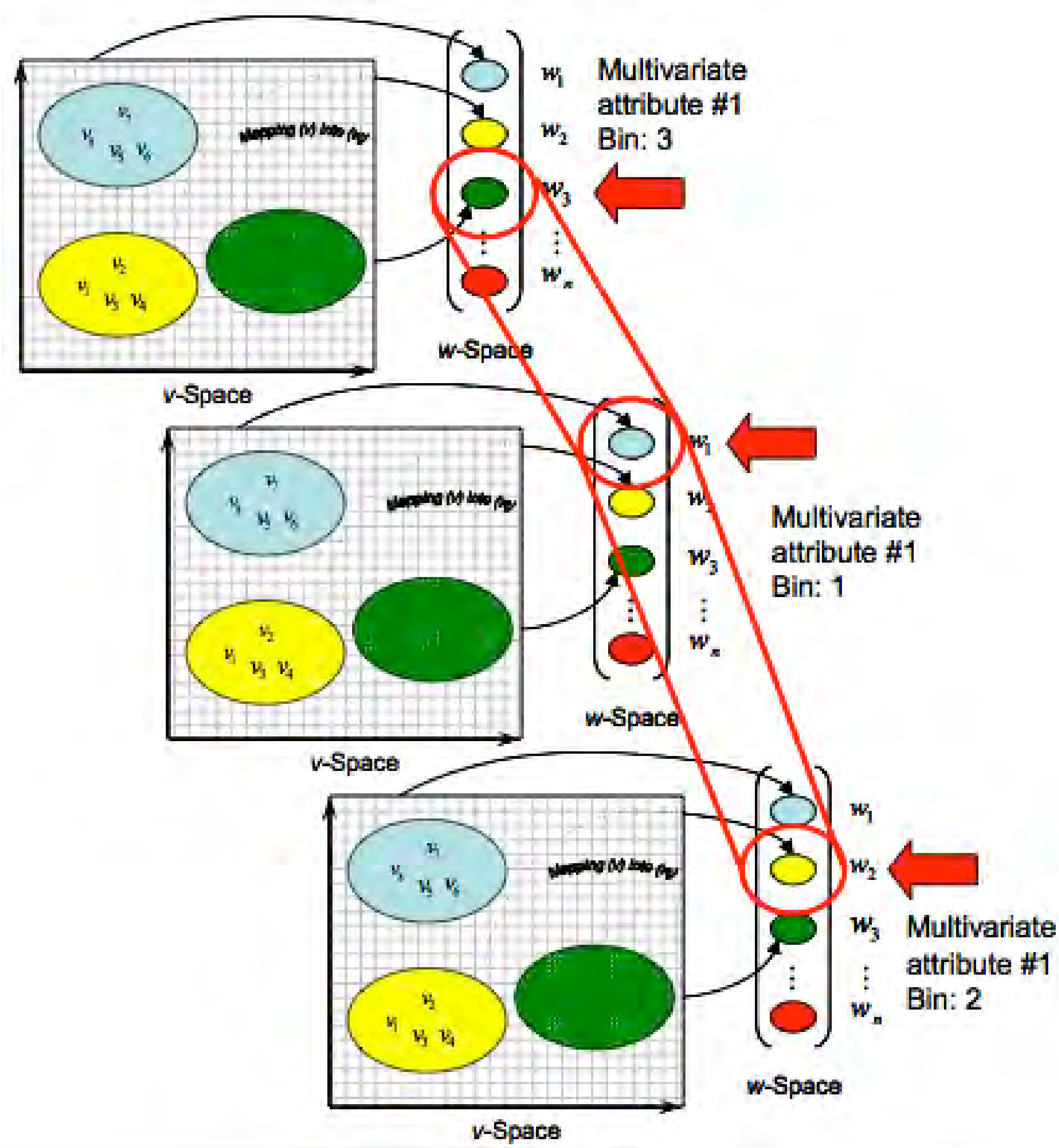


Figure 6. Multivariate segmentation of binned univariate attribute data.

The list of companies or analytic methods is by no means exhaustive, but is meant to supplement the reader's knowledge of capabilities and providers. One interesting observation that could be relevant to social systems theory is that each of these tools are designed to operate against relatively large databases. Some of the vendors specialize in pattern searches through massive amounts of data. While this capability can be extremely useful in detecting patterns at the observation level, the formation of groups (asset classes) should be able to leverage the data reduction methods described in the arithmetic binning section of this work.

One of the complaints of users of commercial data mining tools is the number of segments produced when a search through the entire data repository is conducted. Irrelevant clusters appear and must be merged with other more legitimate business clusters to keep the number to a manageable level. Recent developments in genetic algorithms and other types of linear programming derivatives may give users a necessary control capability with the benefit of fast solution times, especially if the massive amounts of raw data are reduced using arithmetic binning as a pre-processing technique.

Of particular interest in partitioning activities is the use of goal programming (Winston, 1991). Goal programming is used where the decision maker has several objectives and there may be no point in feasible attribute space that satisfies all the objectives at once. Goal programming would provide separable cutting planes by controlling constraint values while solving for objectives in the priority of a weighting scheme where each objective would receive a weighted value (the sum of the weights = 1.0).

The use of goal programming would require the development of distance formulas for minimizing the attribute measures within a cluster, for maximizing the distances between clusters, and separable cluster development based on a financial return vector. These three goals would serve as the shared objective function for the optimization where the constraints may include a minimum numbers of clusters, a minimum number of customers in each cluster, and special handling of small clusters.

Analytic Method	Keywords
Neural Networks	Rule extraction, Machine learning, Cellular automata
Clustering	Segmentation, Partitioning, Kohonen nets, Relational data analysis, K -means, Text mining
Decision Trees	CART, C4.5, Decision rules, Symbolic logic, Rule induction, Predictive models
Regression (Linear, logistic, non -linear)	Bayes, Multivariate, Measures of central tendency, Statistics
Variable selection and dimension reduction	Attribute analysis, Binning, Data reduction, Descritization
Time series analysis	Leading, Lagging, Exponential smoothing, Trend analysis, Seasonal analysis, Box -Jenkins, X -11
Market -basket analysis	Association rules, Sequential patterns, Predictive modeling

Commercial Companies

- Computer Assoc.
- Mathematica
- SAS Institute
- IBM Corp.
- Oracle Corp.
- Neuralware
- HNC
- Angoss
- DataMind
- SPSS
- NCR Corp.
- Mathsoft
- Quadstone
- XLMiner
- GhostMiner
- Genalytics
- Insightful Miner
- KXEN
- Fair Isaac
- Open Source
- Statistica

Table 1. Analytic methods and commercial data mining offerings.

Kinematic Topology: Stability and Transition Matrices

The addition to the segmentation process of the social systems theory concept of kinematics affords the marketing manager an opportunity to view the movements of the social system and specific groups of interest over time in the form of transitions from one cluster to the next. Only user defined segmentation boundaries or constraints would resist change over time. For instance, defining the groups in terms of recency boundaries, frequency boundaries, and monetary value boundaries creates an artificial partitioning that may not reveal the group's sensitivity to changes in the marketplace (Haydock & Bibelnieks, 1999).

Transition matrices would become important kinematic inputs into the asset allocation optimization and show the relationship between prior marketing treatments and payoffs. This feature determines how group members flow from one state to the next as illustrated in Figure 7. At least three periods would be necessary for use in a strategic model in order to forecast the future state which will be a critical optimization input depicting file growth and monetary resource allocation to the various segments (Haydock & Bibelnieks, 1999). Table 2 demonstrates how these combinatorial transitions would look over each time period from any cluster to any other cluster, subject to business rules. Customer flow counts would serve as the business metric.

A key indicator of change or stability is attribution over time. The segmentation was determined by the attributes selected for partitioning. Each of the segments has a set of strong attributes which defined the segments. These strong attributes may be quite different from any one segment to the next.

Understanding the growth or contraction of these strong attribute values over time while utilizing the same time periods as the transition matrices would indicate how the topology is changing relative to the core stable attributes which should be static over the time period. This attribution would become visible in the form of a report and a forecast for the future direction of the social system under investigation.

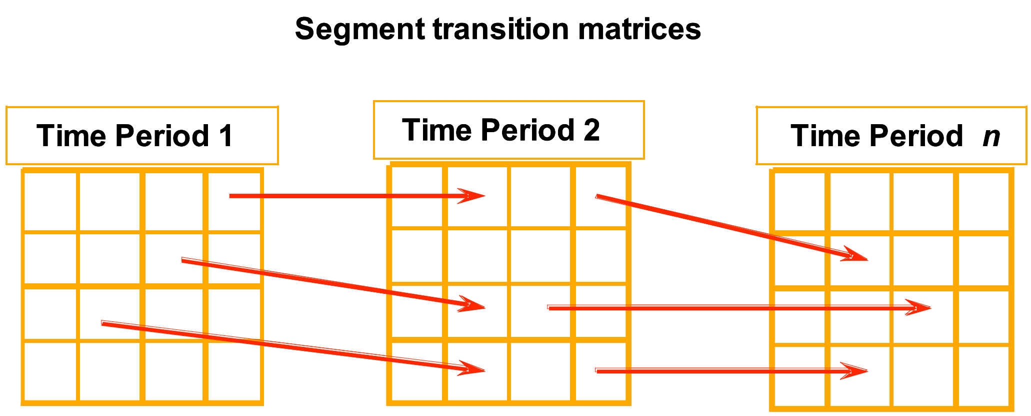


Figure 7. Transition matrices and flows into segments over time.

Transitions of Total Buyers	Year 2003	Year 2004	Year 2005	Year 2006	Year 2007
Single, 0-24 -> Multi, 0-12	2,000	2,200	2,300	2,400	2,500
Single, 25+ -> Multi, 0-12	4,000	4,100	4,200	4,500	4,600
Single, 25+ -> Single, 0-24	2,200	2,400	2,600	2,800	3,000
Multi, 13-24 -> Multi, 0-12	4,000	4,700	4,900	4,500	4,800
Multi, 25-48 -> Multi, 0-12	5,000	5,200	5,400	5,600	5,800
Multi, 49+ -> Multi, 0-12	1,500	1,700	1,900	2,500	3,500

Table 2. Transition matrix showing combinatorial movement over time.

Dynamic State Change: The Probabilistic Future

Systems are at the core of complexity thinking and can be defined as a group of interacting parts functioning as a whole and distinguishable from its surroundings by recognizable boundaries. Bertalanffy (Strauss, 2002) thought that systems have properties that are emergent, that are not intrinsically found within any of the component parts. They exist only at a higher level of description. For instance, customer buying behavior can be captured in the kinetic transition matrices that, from the segment point of view, appear only as flows in and out of clusters. When all the segments are viewed from a total social system point of view the appearance of new properties of the whole become visible as a result of the inter-connection of the segments. The system whole is self-organizing in a dissimilar way from the prior time period.

Cybernetics is a science that is concerned with the treatment of inputs and outputs in machines and biological systems. The majority of systems treated in cybernetics are deterministic. This means that the next state of the system is fully specified by the combination of the system inputs, its current state and the transitions or changes that occur over time (Krivov, Dahiya, & Ashraf, 2002). But another mode is more common in biological and human social systems, and that is the random or stochastic mode. Here the options available are probabilistic, we need to spin the roulette wheel to determine which transitions will be chosen and the quantity of the flow from cluster to cluster. These type of kinematic transitions are indeterminate and operate in a statistically random environment.

Information could be thought of as the opposite of risk. In dealing with the forward movement and the dynamics of this kinematic social system the use of probabilities will help profile the trajectory of the emerging whole system as it will not follow one deterministic path, but instead take one of many paths based on prior states, conditional events, and joint probabilities producing posterior odds of the future. This information will be invaluable to the marketing investor.

The risk inherent in contact economics is the risk of an unprofitable investment in marketing treatments to a target group. Profiling risk occurs when poor investment bets are made based on groups whose past or future buying behaviors are not well understood. Among many investment lessons articulated by Markowitz (1959) is that return dispersion around a probabilistic expectation was one method of diversifying risk. Profiling future kinematic transitions in such a way so as to include several possible states of investment adversity should provide a more stable investment environment for the marketing executive.

This notion gives rise to the statistical philosophy that prior probabilities should incorporate information about things observable. Aczel & Sounderpandian (2002) suggest that Bayes Theorem will allow us to obtain the probability of an event *B* given an event *A* from the probability of event *A* occurring given event *B*. A Bayes formulation also allows for the inclusion of subjective information about the future condition of a group. These unique statistical features give a fundamental measuring device for marketing uncertainty.

The distribution for a parameter in Bayes Theorem is referred to as a prior distribution because it represents the distribution of a degree of belief about the parameter prior to observing the data. The theorem gives a mathematical procedure to update the prior belief about the value of the parameter with data to produce a posterior distribution. For example, the theorem could examine a belief about tomorrow's state of transition about a consumer group, given today's observational data that relate to the behavior and a prior belief about that behavior. This technique could provide key pieces of the puzzle we are trying to solve in explaining and projecting consumer group behavior. Bayes Theorem states that for two events *A* and *B* (Aczel & Sounderpandian, 2002):

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \overline{B})P(\overline{B})}$$

Where:

$P(B \mid A)$ = the probability of event B given event A,

$P(B)$ = the probability of the occurrence of event B.

$P(\overline{B})$ = the probability of the complement of B or, $1 - P(B)$.

Bayes Theorem can also be extended for a vector of events (B_i) where the set is partitioned $B_1, B_2, B_3, ..., B_n$ and represented by (Aczel & Sounderpandian, 2002):

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}.$$

Where:

$P(B_1 \mid A)$ = the probability of one of the sets in the partition B_1 given the occurrence of event A over all events B_i.

The probabilities $P(B)$ and $P(\overline{B})$ are referred to as the prior probabilities of the events B and its complement $\overline{B}$. The probability $P(B \mid A)$ is referred to as the posterior probability of B. In the case that will be described, a partition set will be required utilizing three partitions to describe adversity states.

Relating Bayes theorem to future transitions, and therefore future investment behavior, a notion of adversity is required. Adversity is the state of an endogenous set of events that may work against the success of the marketing investment. Adversity in this context could be classified as minor, major, or zero. The marketing investor wants to understand what customer transition flow will occur and treat this information as an expected value.

As an example, consider a marketing investor merchandising retail apparel and operating in a catalog and internet channel environment. Apparel purchases are directly impacted by economic conditions. Transition counts between the firm's consumer groups can be gathered for each time period. These time periods could be classified as existing in one of three states of adversity (minor, major, and zero). The marketing investor would most likely understand current state conditions which would become the prior distributions (today's state). These conditions would be input as a series of probabilities. Figure 8 illustrates the computation and procedure for obtaining information about future adversity state conditions.

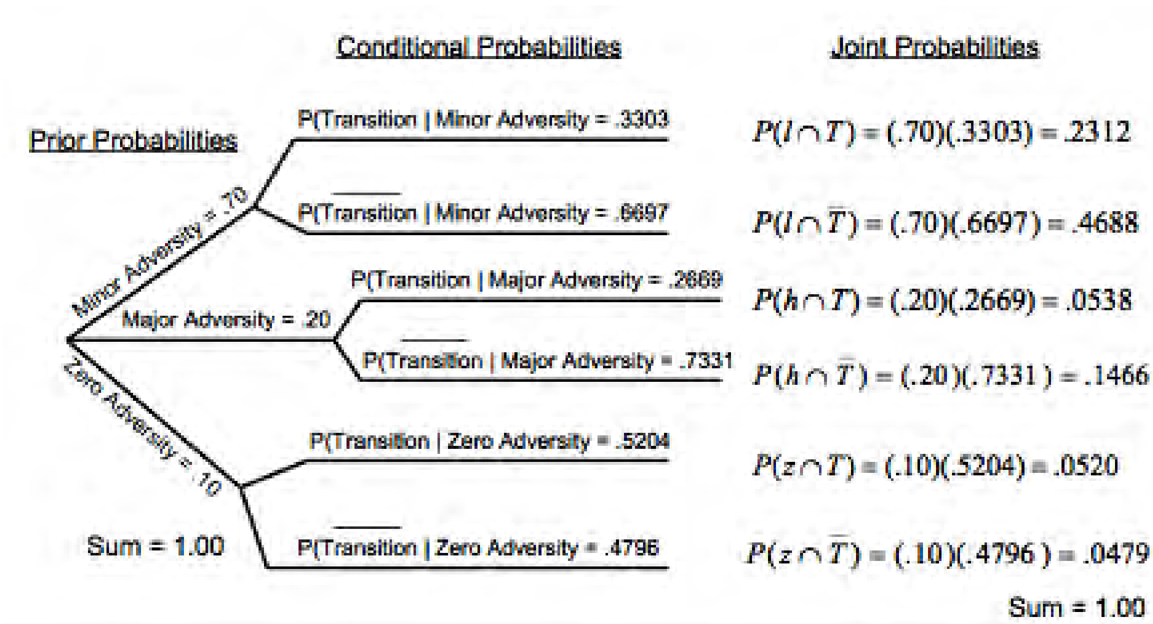


Figure 8. Bayes probability tree to determine adversity state.

Since there are three states of nature, there are three branches to the tree originating from the tree origin. The second component of the tree details the conditional probabilities and the complements of the conditional probabilities. The complements are represented by: $P(\overline{B}) = 1 - P(B)$. The third component of the tree develops the joint probabilities (adversity state x conditional probability).

Figure 9 illustrates the computation of the posterior probabilities. These probabilities formulate the initial conditions for the first time step in the future transitions. They will be used as inputs into the next set of Bayes computations which determines the quantity of transitions under adversity state conditions.

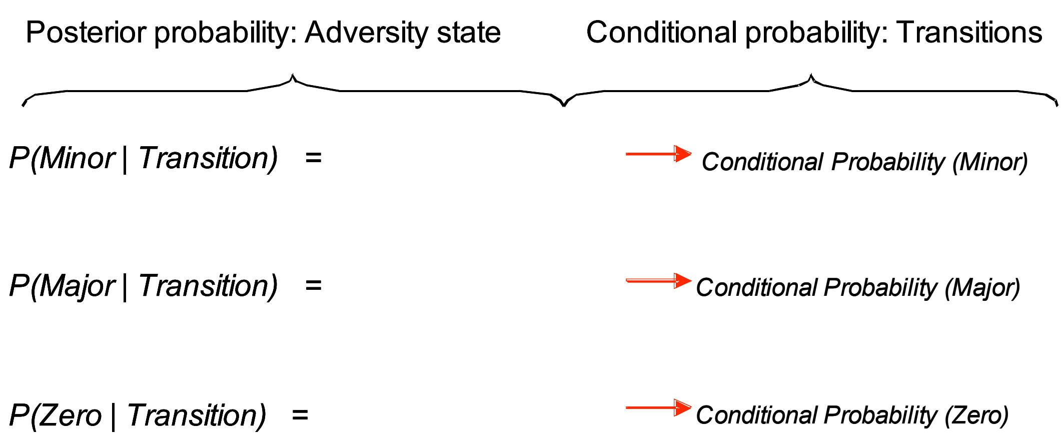


Figure 9. Posterior state probabilities are input into transition conditionals.

The development of the probabilistic transitions is illustrated in Figure 10. Each combinatorial pair of allowable transitions would spawn a set of Bayes posteriors. In Figure 10 a single Bayes probability tree is developed for the transition A→B. From a computation perspective, the tree generation would be iterative based on a table of priors and allowable transitions.

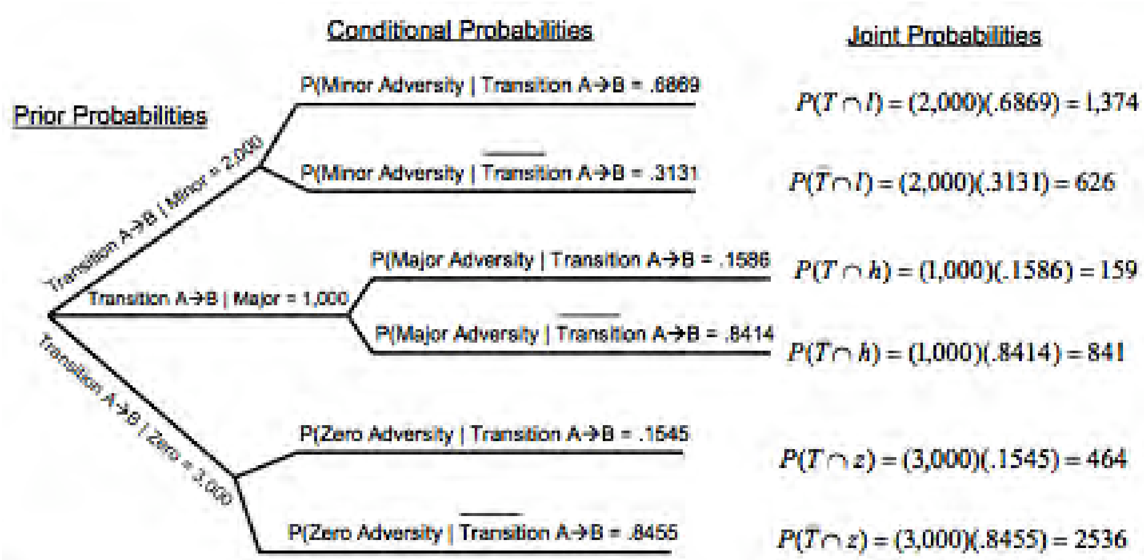


Figure 10. Bayes tree showing transitions under uncertainty from prior states.

Figure 11 illustrates the final steps which are to compute the posterior transition probabilities for transition set A→B and the expected value of the number of transitions under uncertainty.

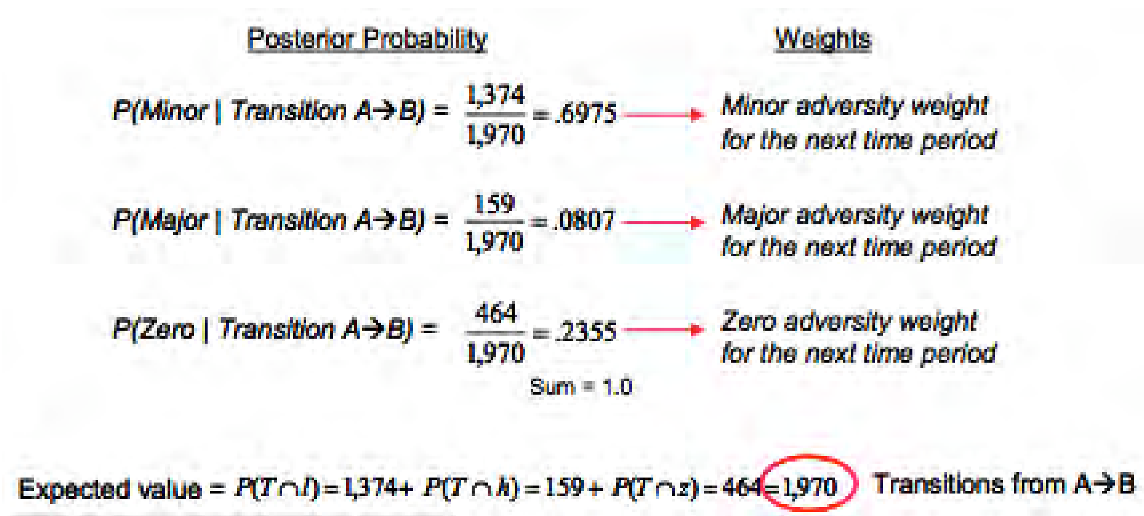


Figure 11. Bayes posterior probabilities and expected value of transitions

The weights generated as a result of the posterior probabilities in the first time step become the weights in the conditional probabilities for the next time step. This process iterates through each of the allowable transitions and for each time step forward until the transition matrix, such as that illustrated in Table 2 is completed. The marketing investor can influence any future state by inputting subjective probabilities into the first Bayes state solution. These probabilistic transitions have captured the uncertainty in the valuation process, which will be detailed next in an articulation of the biotic aspect of the social systems computation.

Biotic Valuation: The Lifetime Value of a Social System

Measuring the lifetime value (LTV) of a segment or group should reflect the buying behavior structure of the industry the model is intended for. Finance, retail, and manufacturing all have different customer strategies, goals and cost structures. The discussion of LTV for these purposes will center on a retail catalog company so that concepts can be illustrated.

The LTV of any individual retail catalog customer could be considered the total amount of cash flows from purchases that would be expected from that customer minus the total amount of costs utilized in servicing that customer for as long as they are expected to be a customer. Essentially what a customer is worth in profit over the long relationship with the firm. These observations would be taken from the purchasing data over a period of time. Most retail catalog companies store a data referred to as “time-as-customer” so the computation from the first time period until the last time period (terminal horizon) is easily constructed from the data.

What are the components then of a typical LTV computation? Faherty (2004) articulates in a report on LTV to a catalog company that the data would have to include behavioral data on both sales and profits generated over the life of the customer. The data would also include the promotion history of the customer, which catalog companies on average do a good job of keeping. This would include any other variable costs obtained by servicing that customer, as well as any revenues generated by servicing the purchases (for instance credit card interest paid to supplement purchases). Some individual data elements to include in the computation may consist of:

- The time of first purchase,
- An expected purchase frequency,
- A loyalty probability transformed into number of time periods as a customer,
- The expected amount of purchases,
- The expectation of using the firms financing alternatives,
- The variable cost of the catalog,
- The average margin for the merchandise category where the purchases were made,
- The expected categories where purchases will be made in the future, and
- Typical merchandise returns.

These factors are just a sample of the types of data which would reside in an equation that would discount the net present value of future cash flows. Each company will have their own customer profit computation so one will not be attempted here. The point is that the value of individual profit measures discounted into LTV metrics is in the measurement space of the groups and the whole social system.

When measuring the LTV of groups, it starts to get interesting, especially when trying to understand the transitions from one group to another group and watching their value increase and decrease with the kinematic movement. Peppers and Rogers (2005) consider the scarce resource in the enterprise equation is the customer, and understanding their value is paramount to a firms success. They calculate a "return on customer" (ROC) to be (Peppers & Rogers, 2005 , p. 7):

$$ROC = \frac{\psi + \Delta CE_i}{CE_{i-1}}.$$

Where: ψ = The cash flows from customers during period *i,*

ΔCE_i = the change in customer equity during period *i,*

CE_{i-1} = the customer equity at the beginning of period *i.*

ROC is equal to a firm's current period cash flow from their customers plus any change in the equity of the customer base, divided by the total customer equity at the beginning of the period (Peppers & Rogers, 2005).

Peppers & Rogers refer to customer lifetime value as the source of consumer equity. If all the lifetime values are summed across all the segments, an equity measure can be derived: the net present value of the future cash flows from the customer base of the firm. This for example, would be nearly the same equation used in valuing a debt instrument of the firm, and would therefore be a familiar concept.

Each segment would measure the inflow and outflow of customers and their transitional valuations. As a customer transits from one segment to another the valuation should change as well reflecting the segment lifetime valuation expectation for all customers in the new group. The sum across the group provides the total customer equity. Peppers and Rogers state that the valuation of the firm should be directly linked to this ROC measure. Biotic valuation is a very powerful decision making tool when added to the segmentation process.

Summary: Taking the Process to the Firm

In this paper the process of segmentation has been proposed to include much more than the usual partition building tasks which create separable groups. The proposed segmentation process would be firmly based on Von Bertalanffy's concept of social systems theory and would include five process components:

- Arithmetic binning,
- Spatial clustering,
- Kinematic topology,
- Dynamic state change, and
- Biotic valuation.

The argument is that deeper knowledge from a management perspective can be gained from understanding the complete social system through the genetic movements of the groups as they mature, age, and transition into a set of diverse self organizing clusters. The management complexity begins where the technical complexity leaves off with the process of socializing the segments and their measurement systems into an enterprise that has most likely been built upon another premise of the customer?

The key is in engaging the firms executives in the process (M.P. Haydock, 2005). The steps may include:

1. Defining the segments and classifying the firm's household file into the segments.
2. Determining the current financial potential of each segment.
3. Engaging the senior executives of the enterprise in selecting the strategic customer segments from the cluster inventory.
4. Identify customer wants and needs and identify groups open to the brand, key purchase triggers, and most likely product they will purchase.
5. Determine the overall brand positioning: which brands to offer, the customer value proposition, and the competitors positions and possible responses.
6. Develop the marketing mix plans for each segment.
7. Design and conduct marketing mix tests per segment using random stratified proportional test and control groups.
8. Revise and adjust segmentation and marketing plans in accordance with what was learned during the tests.
9. Plan the resource investments. Allocate budgets to segments, appoint segment leaders, and market mix investment strategy.

The firms who can implement around the proposed social systems theory concept will have a much better opportunity to gain, maintain, and retain their most valuable asset: their customers.

References

Aczel, A. D., & Sounderpandian, J. (2002). *Complete business statistics* (5th ed.). Boston: McGraw-Hill Irwin.

Borowski, E. J., & Borwein, J. M. (1989). *Collins dictionary of mathematics*. London: Harper Collins.

Faherty, D. (2004). *DII White Paper: Customer Long-term Value (LTV)* (Report). Portland, ME: Decision Intelligence, Inc.

Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems, 13*(2-3), 99-115.

Harsanyi, J. C. (1967). Games with incomplete information played by "Bayesian" players, I-III. *Management Science, 14*(3), 159-182.

Haughton, D., Deichmann, J., Eshghi, A., Sayek, S., Teebagy, N., & Topi, H. (2003). A Review of software packages for data mining. *The American Statistician, 57*(4), 290-309.

Haydock, M. P. (2005). *Target segment process* (Report to a client). Minneapolis: Decision Intelligence, Inc.

Haydock, M. P., & Bibelnieks, E. (Eds.). (1999). *Horizontal marketing: Optimized one-to-one marketing*. London: Macmillan Press.

Hosking, J. R. M., Pednault, E. P. D., & Sudan, M. (1997). A statistical perspective on data mining. *Future Generation Computer Systems, 13*(2-3), 117-134.

Kotler, P. (1994). *Marketing management: Analysis, planning, implementation and control* (8th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Krivov, S., Dahiya, A., & Ashraf, J. (2002). From equations to patters: Logic-based approach to general systems theory. *International Journal of General Systems, 31*(2), 183-205.

Markowitz, H. M. (1959). *Portfolio selection* (2nd ed.). Cambridge, MA: Basil Blackwell, Inc.

Michaud, P. (1997). Clustering techniques. *Future Generation Computer Systems, 13*(2-3), 135-147.

Peppers, D., & Rogers, M. (2005). *Return on customer*. New York: Doubleday.

Roederer, M., Treister, A., Moore, W., & Herzenberg, L. A. (2001). Probability binning comparison: A metric for quantitating univariate distribution differences. *Cytometry, 45*, 37-36.

Samli, A. C., Pohlen, T. L., & Bozovic, N. (2002). A review of data mining techniques as they apply to marketing: Generating strategic information to develop market segments. *Marketing Review, 3*(2), 211-227.

Schurmann, J. (1996). *Pattern Classification: A Unified View of Statistical and Neural Approaches*. New York: John Wiley & Sons, Inc.

Sharpe, W., & Alexander, G. J. (1990). *Investments* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Shepard, D. (2003). Customer segmentation: Problems, solutions. *Direct, 15*(3), 39-40.

Sheth, J. N. (1967). A review of buyer behavior. *Management Science, 13*(12), 718-756.

Strauss, D. F. M. (2002). The scope and limitation of Von Bertalanffy's systems theory. *South African Journal of Philosophy, 21*(3), 163-179.

Winston, W. L. (1991). *Introduction to mathematical programming: Applications and algorithms*. Boston: PWS-Kent Publishing Company.

Zait, M., & Messatfa, H. (1997). A comparative study of clustering methods. *Future Generation Computer Systems, 13*(2-3), 149-159.

Zhang, C., & Cheng, B. (2003). Binning methodology for nonparametric goodness-of-fit test. *Journal of Statistical Computation & Simulation, 73*(1), 71-82.

Implications of Personal Networks (*Ala'aqat*) in the African Context: the Case of Strategy Formation in Two Major Sudanese Enterprises

A. M. Elamin, Ajman University of Science and Technology Network, UAE

ABSTRACT

Personal networks permeate Africa and are considered to be a salient feature of the African culture. The present paper attempts to shed light on the influence of such a vital cultural phenomenon on a managerial process of strategy formation within an enterprise context of an African country (Sudan). There has been an increase in recent years in the number of publications, which focus on strategy formation process from the network perspective by emphasising the totality of inter-organisational relationships direct and indirect, that a firm is involved in. This paper deviates from this view by presenting some strategic implications of viewing strategy formation processes as a function of personal networks instead. Using the grounded approach the paper discovered that there are two types of personal connections that are extensively influence strategy-making processes in the Sudanese context: personal connections between executives at inter-firm level and connections with government officials.

Keywords

Personal connections, Strategy, Grounded approach, Culture

Introduction

This paper reports research done to investigate what processes are used to form strategy in Sudanese enterprises. The prevailing stream in the Western literature accentuates strategy formation processes from a network perspective that emphasising the totality of inter-organisational relationships that a firm is involved in. This paper, however, diverge from this prominent view by presenting some strategic implications of viewing strategy formation processes from personal network perspective.

Objectives of the Research

The objectives of this research paper is to develop a description and explanation of the strategy making processes as a function of personal networks or what we call in Sudanese dialect (***Ala'aqat***) that contribute to a better understanding of the behaviour of the selected Sudanese enterprises when they make important decisions. Moreover, the intention is to generate a guiding framework rather than a comprehensive theory that could help in understanding the strategy making processes in the enterprises, which are subject to similar circumstances.

Literature Review

Perhaps one of the major shortcomings of the contemporary literature on strategy is that it has largely adopted a faceless view of the environment confronting individual firms and has ignored the intra-organizational and inter-organizational relationships at the individual level. Some of the few writings that have addressed the issue have considered it as part of power practices inside and outside the organization. The issue is also superficially touched upon by the growing literature on the network. However, the network perspective emphasizes the totality of inter-organizational relationships direct and indirect, that a firm is involved in, and pays little attention to individuals' relationships.

This section is devoted to elucidating the importance of understanding the relationship between individuals inside and outside organizations on strategy development processes. Many scholars have explained organizational working on the basis of networks (e.g., Granovetter 1973; Carroll and Teo, 1996; Johnson and Scholes 1999; Peng and Luo, 2000). Carroll and Teo, (1996) have identified two types of social networks for managers – 'organizational membership networks' and 'core discussion networks'. Organizational membership networks consist of ties to those organizations to which people formally belong as members. Core discussion networks consist of ties to individuals with whom people discuss important personal matters. Burt states that core discussion networks provide 'a window through which interpersonal environments can be scrutinized' (1992: 317). Carroll and Teo (1996) argue that core discussion networks tell about the strength of ties in various settings and their structure across settings, and thus inform about the relative importance of various social settings to an individual. Carroll and Teo's (1996) division of the social networks of managers into 'organizational membership networks' and 'core discussion networks' seems to be convenient for discussing the relevance of individual relations to the strategy development processes.

Within these two identified types of social networks, managers build managerial ties. Granovetter (1985) defined such ties as an individual's attempt to mobilize personal contacts in order to profit from entrepreneurial opportunities. Powell (1990), on the other hand, defined them at firm level as a firm's efforts to co-operate with others in order to obtain and sustain a competitive advantage. Powell's definition seems very similar to the ideas of co-operative strategy making (Mintzberg et al.1998). In a more comprehensive way, Geletkanycz and Hambrick defined such ties as "executives' boundary-spanning activities and their associated interactions with external entities (1997: 654)". Many scholars suggest that the greater the environmental uncertainty, the more likely it is that firms will rely on managerial ties when entering exchange relationships (Pfeffer and Salancick, 1978; Powell, 1990). Some other scholars have associated the presence of managerial ties with the state of imperfect competition and undeveloped institutional structures. Peng and Luo (2000) argue that the social capital embedded in managerial ties might be more important in imperfect competition characterized by weak institutional support and distorted information. The social capital embedded in managerial ties in transition economies, therefore, might compensate for an inherent lack of market supporting institutions such as transparent laws and regulations associated with such types of economies. Peng and Heath (1996) assert that informal institutional constraints, such as those embodied in the interpersonal ties cultivated by managers, may play a more important role in facilitating economic exchanges and hence assert a more significant impact on firm performance. Taking China as an example, Tsui and Farth (1997) contend that most Chinese people cultivate intricate and pervasive personal ties, called *guanxi* (pronounced " kuanshi"), which govern their attitudes toward social and business relationships. On relationship with government officials, Walder (1995) argues that a network strategy linking newly-founded firms and local officials may lead to better firm performance. Overall, Yeung and Tung emphasize on the importance of personal connections within the developing contexts and state that 'who you know is more important than what you know' (1996:54).

The literature suggests two types of managerial ties. The first are ties with executives at others firm, such as suppliers, buyers and competitors. The second type of tie, which is very important especially in transition economies, is a tie with government officials. In such economies, officials at various levels of the government have considerable power to approve projects and allocate resources (Walder, 1995). Consequently, government intervention remains a constant hazard to many firms (Peng, 1997). Given the need to avoid sources of environmental uncertainty (Pfeffer and Salancik, 1978), managers naturally maintain a 'disproportionately greater contact' with government officials (Child, 1994:154).

On their study of managerial ties in China, Peng and Luo (2000) found that ties with officials are more important than ties with managers at other firms, suggesting that firms may have greater resource dependence on officials than on other firms. This finding is very consistent with Walder's (1995) argument in which he accentuated the power that officials have in a transition economy. Unlike most of the Western-based studies, which emphasize the managers' ties with their peers, Walder (1995) and Peng and Luo (2000) assert that ties with government officials are important in transition economies where the role of the government is still substantial.

The social network provides information, access and support that are not available to those outside it. Consequently these networks could be considered as social resources (Lin et al., 1981; Lai et al., 1990) and social capital (Coleman, 1988). Whittington (1993) argues that people's economic behavior is embedded in a network of social relations that may involve their families, the state, their professional and educational backgrounds, even their religion and ethnicity. These networks influence both the means and ends of action, defining what is appropriate and reasonable behavior for their members. Whittington's idea is consistent with the argument that the strategic choices that managers make inherently reflect their backgrounds and experience (Hambrick and Manson, 1984). In the light of this argument, Peng and Luo state that 'managers' social ties, contacts, and networks are believed to affect firms' strategic choices and performance' (2000: 486). These contacts and networks 'are seen as important conduits for informational and social influences on executive decision making' (Geletkanycz and Hambrick, 1997:655).

Research Methodology

This part presents the methodology utilized in this paper. To explore and understand the influence of personal networks on strategy making processes within the Sudanese context, this author argued for a phenomenological rather than a positivist approach based on grounded theory and case study methods. The ground for such a choice is that the strategic management field is poorly researched in Sudan. In this respect this study represent the first of its kind to explore strategy making from such an angle in that particular context.

Two criteria have been used for selection of the sites for conducting the in-depth investigation. These include suitability and relevance of enterprise for observing strategy making processes and the quality of access achieved. The selected sites are Sudan Telecommunication Limited Company (Sudatel) and the Gum Arabic[1] Limited Company (GAC). The first one is a telecommunications company, which was privatized following the adoption of Sudan's Structural Adjustment Programme in 1992, and the second is an enterprise responsible for international marketing of one of the most important commodities in the country. Both enterprises are organized as public limited companies according to the 'Company Law 1925' in which the government is assumed not to be the prime shareholder.

Two strategic issues within the two selected enterprises have been identified for in-depth investigations. A wide spectrum of research on strategic decision-making has justified a selection of strategic instances for studying strategy processes. Cray et al. (1991) argue that one track for studying strategy processes is to analyze a decision or a series of decisions in a single organization. The works of Allison, 1971; Pettigrew 1973, 1985a; and Mintzberg et al., 1976 contain some examples to the point. Such an approach was also used effectively by the London Business School's SID (strategic investment decisions) research during the 1980s (see Yamamoto 1998: 147).

The steps followed for identification of the strategic issues were as follows: firstly, this researcher contacted the general mangers/managing directors of the selected enterprises and the purpose of the study was presented. Secondly, the contacted persons were asked to nominate issues that had strategic importance to the enterprise and that had recently completed. Although allowing the contacted persons to identify issues ensures both interest and first-hand knowledge (Nutt, 1998), is also not free from pitfalls. In each case the general manager identified a single strategic issue. The contacted person was asked to identify the main actors associated with each selected issue (Hafsi and Hafsi, 1989; Nutt, 1998). The selected issues were:

1. Moving and warehousing Gum Arabic in Dubai free zone area for depositing and exporting (Gum Arabic limited Company 'GAC').
2. The introduction of mobile services in the Greater Khartoum and the establishment of new mobile telecommunication company (Mobitel) to provide this service (Sudan Telecommunication limited Company 'Sudatel').

Data collection relied heavily on the in-depth interviewing. Interviews carried out by this researcher should be categorised into a semi-structured one. Interview guides were prepared before each interview took place. Each interview's questions were based on feedback from the previous interviews. The focus of the interviews was on the conditions that stimulate the identified strategic issue, actions/interactions of the key players and the consequences of these responses. The emphasis of this researcher's intention was to discover the processes through which these strategic issues were carried out. During interviewing, this researcher was maintaining a balance between excessive passivity and over-direction. Also, data obtained from interviews were cross-checked with interviewees against each other, and against the documentary evidence (i.e. triangulation). The people interviewed include chairmen of the board, board members, general managers, relevant departmental managers, and top government officials in relevant ministries. Informal discussion was also held with some other concerned

1 Gum Arabic is a trade name for a natural, organic forest product from the genus Acacia. Sudan monopolises its production by producing 85 to 90% of the total annual world production. Gum Arabic represents one of the main Sudanese export products.

people outside the enterprises, 'consultants, and academicians'. Thirty-eight interviews were conducted resulting in forty-nine tape-recorded hours. Alongside the primary data, considerable materials were collected about the issues, enterprises and the broader context of the study. The main source of these materials was the enterprises' files and official reports.

For data analysis, this researcher uses 'grounded theory analytical procedures' (Strauss and Corbin, 1998) supported by the 'within case and across case analysis' recommended by Eisenhardt (1989) and 'replication logic' suggested by Yin, (1989).

Elaboration

The phases of recognition of problem, search for solution, selection of solution, and implementation are basic parts of strategy development processes within the two Sudanese enterprises studied. The existence or otherwise of the authorization as a separate phase within strategy development processes depends on the existence of government at a supervisory level of enterprise. This study found the authorization existed as a separate phase when the government appeared at the supervisory level of an enterprise namely in the case of GAC. Several studies accentuated the role of government as an influential external player in the authorization of strategic decisions within enterprises (Mintzberg et al., 1976), specially in African contexts (Miller et al., 1998), where this study was conducted.

"*Ala'aqat*" refers to existing personal relationships and connections between individuals within an enterprise, and between them and individuals outside. Strategy development processes in the Sudanese enterprises studied extensively drew on personal relationships and connections. Such personal relationships and connections are based on friendship built on attendance at the same educational institution, professional association, and affiliation to the same political party.

Gum Arabic Limited Company (GAC) Case:

The present management of the GAC has justified the strategic decision of moving gum Arabic to Dubai free zone area for depositing and exporting on the ground of the severe financial problem attacked the company in mid 1990s. The recognition of such a financial problem was delayed for two years because the personal relationship between the former general manager and the minister of Foreign Trade (whose the GAC is considered under technical supervision of his ministry) was not cordial. The recognition of the problem took place after a new general manager, who had maintained a cordial relationship with such a minister, replaced the former general manager. The new general manager is a classmate of the minister during university and both were belonging to a single political party. This point indicates that the quality of personal relationships between the top people inside an enterprise, in this case the former general manager, and outside the enterprise, in this case the minister of Foreign Trade, is a crucial aspect for progress of any particular issue. Another point about recognition of the GAC's financial problem was that when the conflict turned out to be severe between the minister of Foreign Trade and the former general manager, both of them resorted to their personal relationships outside the GAC, and its supervisory ministry. The general manager contacted personally the president of the country, whereas the minister contacted the leader of the political party supporting the government. When the leader of the political party intervened and resolved the conflict, he took into consideration the complexity and sophistication of personal relationships among the conflicting persons, and their supporters outside the GAC and its supervisory ministry.

Moreover, to isolate the chairman and his supporters within the BoDs of the GAC from taking part in the definition of the GAC's problem, the new general manager relied on his personal relationships with the minister of Foreign Trade to the extent that with the consent of such a minister, the general manager introduced a policy that restricted the participation of the BoDs. Furthermore, during the authorization process of the decision, the general manager relied heavily on his personal relationships with the governor of the Bank of Sudan 'BS'(whose responsible for approving export operations) and the minister of Foreign Trade for securing their consent on the issue. Similarity in education level, profession, political party membership and attitudes towards interest strengthened the personal relationships between the governor of the BS and the general manager. The instant approval of minister of Foreign Trade for the issue was based on the agreement achieved between the governor of the BS and the general manager. Furthermore, during the implementation phase of the decision, the minister, on the recommendation of the general manager replaced the opponents to the decision issue. Such a decision was based on a personal relationship between the minister and the general manager. The above-discussed type of personal relationships could be categorized as 'relationships with top government officials'.

There is an other type of personal relationships, which I refer to here as 'relationships with executives in other organizations'. The organization could be a supplier, a financial institution, customer, public etc. This type of relationship appeared during the search phase where the GAC management was searching for solutions to the GAC's acute financial problem. In that respect, the success of the general manager in obtaining depositors within a less-familiar source (Dubai) was based on the networks that he had developed during his former employment with Fisal Islamic Bank. Another example of this type of relationship was the relationship between the chairman of BoDs and his supporters, and the GAC's external agents. Maintenance of the competitive position of the GAC was to a great extent based on such types of relationship as opposed to those with top government officials.

Sudan Telecommunication Limited Company (Sudatel) case:

The Sudatel case is very similar to the GAC regards the *Ala'aqat* aspect. Two groups exist within Sudatel executive committee who have substantial influence on decision of introducing the mobile into the greater Khartoum and establishing an independent mobile company. The government group whose members have been appointed by the government and its supporting political party on one hand. On the other hand, the group represents the private foreign and indigenous investors in Sudatel. The government group within the executive committee made a political move by establishing the technical committee to be used as a vehicle to challenge Nepostel's1 plan for introducing the mobile system. The Nepostel's plan was supported by the group represents the private foreign and indigenous investors in Sudatel The appointment of the technical committee members was based on previous personal relationships between the appointed members and the government's group within the executive committee. The fact that the appointed members of the committee were also members of the political party supporting the government indicates the existence of former personal relationships between these people at the political party level. In a similar vein, the appointment of the nuclear management team for the mobile project was also based on former personal relationships between the appointed members and some members within the executive committee. Furthermore, the appearance of the Nile Water Company (NWC) in scene as a competitor to the Daewoo

1 Nepostel: a Dutch company recruited to provide management and consultancy services for Sudatel in 1993.

Corporation, and consideration of its offer to be a Sudatel partner to operate the new mobile company, was based on the personal relationships between the NWC representative and the members of government group within the executive committee. Similarly, the lenient attitude by the government group within the executive committee towards the NWC was fundamentally based on the personal relationships between members of government group and the chairman of the BoDs of the NWC. Also, for the same reason, the government group within the executive committee violated the criteria identified for the selection of the partner to operate the mobile company with Sudatel. Consequently, based on consideration of personal relations, the government group within the executive committee rejected Daewoo's offer and accepted the NWC's one. Based on non-existence of personal relations the executive committee also rejected El Munsory and National Insurance Fund offers' although they were identical to the offer of the NWC's especially with regard to the dissatisfaction of the criterion of technical know-how. Moreover, based on personal relationships, the chairman of the Sudatel's BoDs contacted the top government official to intervene and to protect the contract signed with the NWC against Daewoo's appeal to the minister of transport and telecommunication.

Findings

The elaboration of *Ala'aqat* concept and its influence on strategy development processes within the Sudanese enterprises studied revealed several implications.

First, the existence of *Ala'aqat* as an important influential factor on strategy development processes within the GAC and Sudatel could be considered as part and parcel of the importance of personal connections all over the continent of Africa. Such a phenomenon permeates African social systems and influences the African way of life. The influence of such a phenomenon on African management was well addressed by Blunt and Jones who point out that:

"African managers are linked to extensive networks of social obligations; and that values of reciprocity and social exchange that characterize African social systems still impinge directly on the working lives of African managers" (1992:37).

In a similar vein, Nnadozie considered personal connections as a key to successful business in the continent of Africa by stating that:

"A lot of importance is placed on personal connections and relationships. Africans emphasize knowing people, having personal connections, and relationships. Hence, personal relationships and connections are key to successful business in Africa" (2001:56).

Second, two types of *Ala'aqat* appeared to be critical for strategy development processes within the Sudanese enterprises studied. The first is *Ala'aqat* and its quality between the top person (s) at the enterprise level, and between them and the top government officials and political party's leaders. The second is *Ala'aqat* with executives in other organizations (suppliers, financial institutions, customers, public etc). Both types of *Ala'aqat* identified within the Sudanese enterprises studied had found an early expression in Mintzberg's (1973) early ideas on the *Nature of Managerial Work*. Mintzberg called attention to the liaison role of managers, in which he argued that the chief executives had contacts with wide range of people including subordinates, clients, business associates, suppliers, managers of similar organizations, government and trade union officials, and so on. Most of these contacts were used primarily for gathering information. However, within the context of the GAC and Sudatel, the function of *Ala'aqat* not only restricted to the informational role, but also extended to influence the actual processes of decision-making. Such a difference could be attributed to the contextual differences between the Western milieu where Mintzberg's study was conducted and the Sudanese one.

Both types of *Ala'aqat* appeared to be critical during the whole process of strategy making within the Sudanese enterprises studied. This finding is consistent with a wide stream of studies (Pfeffer and Salancick, 1978; Powell, 1990; Geletkanycz and Hambrick, 1997; Peng and Luo, 2000; Park and Luo, 2001) that emphasized the importance of managerial ties for securing against environmental uncertainty, and facilitating strategic choice. Moreover, the first type of *Ala'aqat* appeared to be more critical during the selection and authorization phases of strategy making process, and during resolving conflict and disputes that are likely to occur between top people. Peng and Luo (2000) and Child (1994) identified a similar type of such relationship within the Chinese context. Moreover, Peng and Luo (2000) found that the managerial ties with officials are more important than ties with managers at other firms. Given the importance of such ties with officials in the Sudanese enterprises studied, however, the influence of such relations across the Sudanese and Chinese contexts was not identical. While it positively influenced strategic development of the Chinese enterprises (Luo and Chen, 1997; Qi, 2000), it appeared to adversely influence the strategic development of Sudanese ones. Apparently, the reason is twofold. First, the fast pace of growth of Chinese economy compared to the slow Sudanese one. Second, the totalitarian political system of Sudan gave the political party supporting the government wide scope in directing such relations for realizing both personal and party agendas rather than the enterprises' strategic development objectives.

The second type of *Ala'aqat* appeared very crucial during the search and selection phases. Several studies (Geletkanycz and Hambrick, 1997; Peng and Luo 2000) have supported such a finding. For instance, Geletkanycz and Hambrick, (1997) emphasized the informational and social influences of managerial ties based on social networks on helping managers to cope with uncertainty inherent in the choice process. They argue that:

"external referents offer models that expand the range of strategic options available for selection" (1997:456).

Similarly, the finding is also supported by Peng and Luo (2000) who emphasized the importance of a manager's social ties in strategy-making process and choice. They argued persuasively that:

"managers' social ties, contacts and networks are believed to effect firms strategic choices and performance" (2000: 486).

Third, the bases for such two types of *Ala'aqat* were found to have stemmed from the educational, professional, political, and most importantly the attitudes towards the issue of interest shared by the members of these networks. The members involved in theses two types of *Ala'aqat* were found either to have attended a similar university or college, or affiliated to a single professional association or most likely affiliated to a single political party (Sudanese National Islamic Front). This finding concurred with those of Carroll and Teo who found that:

"An efficient way for a manager to establish an external network of personal ties is through formal membership in various clubs and societies… involvement in clubs and societies allows managers to get to know others with similar social interests, political affiliations, educational backgrounds and professional work experiences" (2000: 425).

Fourth, *Ala'aqat* with executives in other organizations appeared to be more critical for strategic development of the Sudanese enterprises studied than *Ala'aqat* between the top person(s) at the enterprise level, and between them and the top government officials and political party leaders. This finding seems consistent with Granovetter's (1973) concept of "strong ties" and "weak ties", and Carroll and Teo's (1996) categories of "organizational membership networks" and "core discussion networks". On the one hand, *Ala'aqat* with executives in other organizations resemble both Granovetter's (1973) concept of "weak ties", and Carroll and Teo's (1996) category of "organizational membership networks". On the other, *Ala'aqat* with top government officials and between the top person(s) at the enterprise level, and between them and the top government officials and political party's leader resemble both Granovetter's (1973) concept of "strong ties" and Carroll and Teo's (1996) category of "core discussion networks". The ensuing analysis proved that in both Sudanese enterprises studied, strong emphasis had been put on relationships within top officials of government institutions and its supporting political party's leaders, which represent "strong ties" within managers' "core discussion networks", while little emphasis had been put on relationships with customers, suppliers and external agents, which represent "weak ties" that extended over "organizational membership networks". For preventing environmental uncertainty and taking advantages of opportunities, an organization needs wide contacts with the external environment, which could better be secured through weak ties than strong ones. Granovetter states that:

"Weak ties are more likely to link members of different small groups than are strong ones, which tend to be concentrated within particular groups" (1973: 1376).

In a similar vein, Carroll and Teo (1996) contend that organizational membership networks could connect a manager to a broad range of influential individuals and organizations.

Fifth, the findings of this study concur with the widely held Western view (Mintzberg, 1973; Geletkanycz and Hambrick, 1997) that emphasized the informational role of *Ala'aqat*, and cast attention to the *Ala'aqat's* actual role in strategy-making processes. The last point was supported by Peng and Luo (2000); Qi (2000); Park and Luo, (2001), who found that *Ala'aqat* influences the actual processes of strategy-making especially in a state of transition economy and imperfect competition characterized by weak institutional support and distorted information. The above-cited studies that accentuated the actual role of *Ala'aqat* in strategy-making processes were conducted within the Chinese context, which could represent a good example of a transition economy. Peng and Luo cogently comment that:

"In an environment where formal institutional constraints as laws and regulations are weak, informal institutional constraints, such as those embodied in the interpersonal ties cultivated by managers (for instance, *blatin* in Russia and *guanxi* in China), may play a more important role in facilitating economic exchanges and hence assert a more significant impact on firm performance" (2000:487).

Despite the wide-ranging differences between the Sudanese context and the Chinese one, two common features permeate these countries; namely, the existence of "institutional voids" because of a lack of market-supporting institutions (Peng and Luo, 2000), and the existence of a collectivist culture which reinforced the importance of networks of interpersonal relations. In the Sudan the collectivist culture is fostered by the tribalistic nature of Sudanese society, whereas in China it widely stems from Confucianism (Park and Luo, 2001).

Summary

The existence of *Ala'aqat* is an important influential factor on strategy development processes within the Sudanese enterprises studied; it could be considered as part and parcel of the importance of personal connections all over the continent of Africa. Strategy development within both the GAC and Sudatel is influenced by two types of *Ala'aqat.* The first is *Ala'aqat* and its quality between the top person (s) at the enterprise level, and between them and the top government officials and political party leaders. The second is *Ala'aqat* with executives in other organizations (suppliers, financial institutions, customers, public…etc). The base of both types of *Ala'aqat* stemmed from a similar educational, professional, and political background shared by the members who constitutes such a relation. Both the presence, and influence of both types of *Ala'aqat* on strategy development processes is fairly supported by the existing literature within the strategy field. This study uncovered that *Ala'aqat* function is not only restricted to provision of information as is widely predominant within Western contexts. *Ala'aqat* was found to play a critical role in strategy development processes within the Sudanese enterprises studied. Some, studies based on non-Western contexts that are undergoing transitional economy (China as an example), supported such findings. There were common grounds between the Chinese context and the Sudanese one including 'institutional voids' and collectivist culture. Although this comparison is only restricted to the Sudan and China, I speculate that it would be valid for countries all over the globe that are undergoing similar circumstances. While *Ala'aqat* with government officials positively influenced the strategic development of Chinese firms, it adversely influenced the strategic development of the Sudanese enterprises studied. Based on such contradictions the study showed that the first type of *Ala'aqat's* influences on strategic development of the enterprises studied is subject to the economic conditions and political philosophy followed in the country. The study also found that for the strategic development of the Sudanese enterprises studied, strong emphasis should be put on the second type of *Ala'aqat.*

References

Allison, G. T. (1971) 'Essence of decision: explaining the Cuban Missile crisis', Boston, MA: Little, Brown and Company.

Blunt, P. and Jones, M. L. (1992) 'Managing organisations in Africa', Walter de Gryter and Co.

Burt, R. S. (1980) 'Cooptive corporate actor networks: a consideration of interlocking directorates involving American manufacturing', *Administrative Science Quarterly* 25: 557-582.

Burt, R. S. (1992) 'Structural holes', Cambridge: Harvard University Press.

Carroll, G., and Teo, A. (1996) 'On the social networks of managers' *Academy of Management Journal* 39: 2, 421-440.

Carroll, G., J. Goodstein, and A. Gyenes (1990) 'Managing the institutional environment: evidence from Hungarian agricultural co-operatives', *European Sociological Review* 6: 73-86.

Child, J. (1994) 'Management in China during the age of reform', Cambridge, England: Cambridge University Press.

Child, J. and Smith, C. (1987) 'The context and process of organisational transformation - Cadbury limited in its sector', *Journal of Management Studies* 24: 565-593.

Coleman, J. S. (1988) 'Social capital in the creation of human capital', *American Journal of Sociology* 94: S95- S120.

Cray, G. R., G. R. Mallory, R. J. Butler, Hickson, D. J., and D. C. Wilson (1991) ‘Explaining decision processes’, *Journal of Management Studies* 28(3): 227-251.

Cyert, R. M., and March, J. C. (1963) ‘A behavioural theory of the firm’, Englewood Cliffs, NJ: Prentice-Hall.

Eisenhardt, K. M. (1989) ‘Building theories from case study research’, *Academy of Management Review* 14: 532-550.

Fischer, C. S. (1982) ‘To dwell among friends: personal networks in town and city’, Chicago: University of Chicago Press.

Geletkanycz, M. A., and Hambrick, D. C. (1997) ‘The external ties of top executives: implications for strategic choices and performance’, *Administrative Science Quarterly* 42: 654-681.

Granovetter, M. S. (1973) ‘The strength of weak ties’, *American Journal of Sociology* 78: 1360-1380.

Granovetter, M. S. (1985) ‘Economic action and social structure: the problem of embeddedness’, *American Journal of Sociology* 91: 193-206.

Hafsi, T. and Hafsi, C. (1989) ‘Business behaviour of the Algerian state-controlled enterprises: a decision making process approach’, in T. Hafsi, eds., *Strategic issues in state-controlled enterprises*, JAI Press Inc., London.

Hambrick, D. C. and Mason, P. A. (1984) ‘Upper echelons: the organisation as a reflection of its top managers’, *Academy of Management Review* 9: 193- 206.

Hickson, D. J., R. J. Butler, D. Cray, G. R. Mallory and D. C. Wilson (1986) ‘Top decisions: strategic decision-making in organisation’, Oxford: Basil Blakwell. San Francisco: Jossey-Bass.

Johnson, G., and Scholes, K. (1999) ‘Exploring corporate strategy – text and cases’, (5th ed), Herts,: Prentice Hall.

Kotter, J.P. (1982) ‘ The general managers’, New York: Free Press.

Krackhardt, D. (1990) ‘Assessing the political landscape: structure, cognition and power in organisations’, *Administrative Science Quarterly* 35: 342-369.

Lai, G.W., S. Leung, and N. Lin (1990) ‘Network resources, contact resources, and status attainment: structural and action effects of social resources’, *Paper presented at the annual meeting of the American Sociological association*, Washington, DC.

Lin, N., W. M. Ensel, and J. C. Vaughn (1981) ‘Social resources and strength of ties: Structural factors in occupational status attainment’, *American Sociological Review* 46: 393-405.

Lincoln, J. R. and Miller, J. (1979) ‘Work and friendship ties in organisations, *Administrative Science Quarterly* 24: 181-199.

Lu, Y. (1998) ‘Strategic investment decisions in China’ in V. Papadakis and P. Barwise, eds., *Strategic Decisions*, Kluwer Academic Publishers, the Netherlands.

Luo, Y., and Chen, M. (1997) ‘Does guanxi influence firm performance?’, *Asia Pacific Journal of Management* 14: 1-16.

Marsden, P. V. (1987) ‘Core discussion networks of American’, *American Sociological Review* 52: 122-131.

Miller, S., D. Hickson, and D. Wilson (1998) ‘Decision-Making in Organisations’, Chapter 2.3 in Clegg, S. R., et al., eds., *Handbook of Organisation Studies*, London: Sage.

Mintzberg, H. (1973) ‘The nature of managerial work’, New York: Harper and Row.

Mintzberg, H., D. Raisinghani and A. Theoret (1976) ‘The structure of the “unstructured” decision processes’ *Administrative Science Quarterly* 21: 246-275.

Mintzberg, H., J. Lampel and B. Ahlstrand (1998) ‘ Strategy Safari: A Guided tour through the wilds of strategic management’ Printice Hall.

Mizruchi, M. S. and Stearns, L. B. (1988) ‘A longitudinal study of the formation of interlocking directorates’, *Administrative Science Quarterly* 33:194-210.

Nutt, P. C. (1984) ‘Types of organisational decision processes’, *Administrative Science Quarterly*, 29, 414- 450.

Nutt, P. C. (1998) ‘Framing strategic decisions’, *Organisation Science* 9(2): 195-216.

Palmer, D. P., D. Jennings and X. Zhou (1993) ‘Late adoption of multidivisional form by large U.S. corporation: institutional, political and economic accounts’, *Administrative Science Quarterly* 38: 100-131.

Papadakis, V., S. Lioukas, and D. Chambers (1998) ‘Strategic decision-making processes: the role of management and context’, *Strategic Management Journal* 19: 115-147.

Park, S. H. and Luo, Y. (2001) ‘Guanxi and organisational dynamics: organisational networking in Chinese firms’, *Strategic Management Journal* 22: 455-477.

Peng, M. W. (1997) ‘Firm growth in transitional economy: three longitudinal cases from China, 1989-96, *Organisation Studies* 18: 385-413.

Peng, M. W. and Heath, P. S. (1996) ‘The growth of the firm in planned economies in transition: institutions, organisation and strategic choice’, *Academy of Management Review* 21: 492-528.

Peng, M. W. and Luo, Y. (2000) ‘Managerial ties and firm performance in a transition economy: the nature of a micro-macro link’, *Academy of Management Journal* 43(3): 486-501

Pettigrew, A.M. (1973) ‘The politics of organisational decision making’, Tavistock, London.

Pettigrew, A.M. (1985a) ‘The awakening giant: continuity and change in Imperial Chemical Industries’, Oxford, Basil Blackwell.

Pfeffer, J. (1972) ‘Size and composition of corporate boards of directors: the organisation and its environment’, *Administrative Science Quarterly* 17: 218-229.

Pfeffer, J. (1974) ‘Cooptation and the composition of electric utility boards of directors’, *Pacific Sociological Review* 17: 333-363.

Pfeffer, J. and Salancik G. R. (1978) ‘The external control of organisations: a resource dependence perspective, New York: Harper and Row.

Powell, W. W. (1990) 'Neither market nor hierarchy: network forms of organisation' in B.M. Staw and L. L. Cummings, eds., *Research in Organisational Behaviour*, vol. 12: 295-336, Greenwich, CT: JAI Press.

Qi, H. (2000) 'Exploring the strategic development of Chinese township and village enterprises', *Unpublished Ph.D. thesis*, Lancaster University.

Strauss, A. L. and Corbin, J. (1998) 'Basics of qualitative research: techniques and procedures for developing grounded theory', 2nd edition, Thousand Oaks, CA: Sage.

Sudanow (1992) 'Sudan's structural adjustment policies: A step forward', An English language magazine published monthly by the Ministry of Information and Culture, pp. 22-26, Sept.

The Company Law 1925, Khartoum, Sudan.

The three-year economic salvation programme (1990-1993), Ministry of Finance and economic planning, Khartoum, Sudan.

Thompson, J. (1967) 'Organisations in action', New York: McGraw-Hill.

Tsui, A. S. and Farh, J.- L. (1997) 'Where guanxi matters: relational demography in the Chinese context, *Work and Occupations* 24: 56-79.

Walder, A. G. (1995) 'Local governments as industrial firms: an organisational analysis of China's transitional economy', *American Journal of Sociology* 101: 263-301.

Walsham, G. (1995) 'Interpretative case studies in IS research: nature and method', *European Journal of Information Systems* 4: 74- 81.

Whittington, R. (1993) 'What is strategy – and does it matter?', Routledge, London.

Wright, E. O. and Cho, D. (1992) 'The relative permeability of class boundaries to cross-class friendships: a comparison of the United States, Canada, Sweden and Norway', *American Sociological Review* 57: 85-102.

Yamamoto, M. (1998) 'Strategic decisions and corporate governance in Japan', in V. Papadakis and P. Barwise, eds., *Strategic Decisions*, Kluwer Academic Publishers, the Netherlands.

Yeung, I., and Tung, R. L. (1996) 'Achieving business success in Confucian societies: the importance of guanxi (connections)', *Organisational Dynamics* 25(2): 54-65.

Yin, R. K. (1989) 'Case study research: design and methods', revised edition, Sage publications, Inc.

Zald, M. (1969) 'The power and function of Broad of directors: a theoretical synthesis', *American Journal of Sociology* 75: 97-111.

Predicting Next Page Access by Markov Models and Association Rules on Web Log Data

Siriporn Chimphlee, Suan Dusit Rajabhat University, Thailand
Naomie Salim, University Technology of Malaysia
Mohd Salihin Bin Ngadiman, University Technology of Malaysia
Witcha Chimphlee, Suan Dusit Rajabhat University, Thailand
Surat Srinoy, Suan Dusit Rajabhat University, Thailand

Abstract

Mining user patterns of log file can provide significant and useful informative knowledge. A large amount of the research has been concentrated on trying to correctly predict the pages a user will request. This task requires the development of models that can predict a user's next request to a web server. In this paper, we propose a method for constructing first-order and second-order Markov models of Web site based on past visitor behavior compare with association rules technique. This algorithm has been used to cluster Web site with similar transition behaviors and compares the transition matrix to an optimal size for efficient used to further improve the efficiency of prediction. From this comparison we propose a best overall method and empirically test the proposed model on real web logs.

Keywords

Web mining, Markov Model, Association Rule, Prediction

1. Introduction

The rapid expansion of the World Wide Web has created an unprecedented opportunity to disseminate and gather information online. There is an increasing need to study web-user behavior to better serve the web users and increase the value of enterprises. One important data source for this study is the web-server log data that traces the user's web browsing actions. The web log data consists of sequences of URLs requested by different clients bearing different IP Addresses. Association rules can be used to decide the next likely web page requests based on significant statistical correlations. The result of accurate prediction can be used for recommending products to the customers, suggesting useful links, as well as pre-sending, pre-fetching and caching of web pages for reducing access latency (Yang et al., 2004). The work by Liu et al. (1998) and Wang et al. (2000) considered using association rules for prediction by selecting rules based on confidence measures, but they did not consider the classifiers for sequential data (Yang et al., 2004). It has been observed that users tend to repeat the trails they have followed once (Pitkow and Pirolli, 1999). The better prediction of a user's next request could be made on the data pertaining to that particular user, not all the users. However, this would require reliable user identification and tracking users between sessions. This is usually achieved by sending cookies to a client browser, or by registering users. Both require user cooperation and might discourage some of potential site visitors. So many web sites choose not to use these means of user tracking. Also, building prediction models on individual data would require that users have accessed enough pages to make a prediction, which is not usually the case for a university website that has many casual users (Chakoula, 2000).

In the network system area, Markov chain models have been proposed for capturing browsing paths that occur frequently (Pitkow and Pirolli, 1999; Su et al., 2000). However, researchers in this area did not study the prediction models in the context of association rules, and they did not perform any comparison with other potential prediction models in a systematic way. As a result, it remains an open question how to construct the best association rule based prediction models for web log data (Yang et al., 2004).

This paper is organized as follows. In section 2, we discuss the background and review the past works in related research. In section 3, we present the experimental design. In section 4, we discuss the experimental result. We conclude our work in section 5.

2. Background of study

Nowadays, there are many commercial web site log analysis programmed that can summarize statistics, number of hits common into web site or overview of web page hits or accesses. It is useful information, but none progress toward to understanding of behavior of web users. This study can significant the behavior of web users and more specific draw the behavior based on discover models of their web u sage data, and use to predict the next access with high accuracy.

2.1 Web Mining

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services (Etzioni, 1996). The web mining system uses information determined from the history of the investigated web system. When valuable hidden knowledge about the system of interest has been discovered, this information can be incorporated into a decision support system to improve the performance of the system. Three major web mining methods are web content mining, web structure mining and web usage mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining aims to generate structural summaries about web sites and web pages (Madria and Bhowmick, 1999). Web usage mining is to discover usage patterns from web data, in order to understand and better serve the needs of web-based application. It is an essential step to understand the users' navigation preferences in the study of the quality of an electronic commerce site. In fact, understanding the most likely access patterns of users allows the service provider to personalize and adapt the site's interface for the individual user, and to improve the site's structure.

2.2 Association rules

Association rules (Agraval and Srikant, 1994) were proposed to capture the co-occurrence of buying different items in a supermarket shopping. Association rule generation can be used to relate pages that are most often referenced together in a single server session (Srivastava et al., 2000). In the context of Web usage mining, association rules refer to set of pages that are accessed together with a support value exceeding some specified threshold. The association rules may also serves as a heuristic for prefetching documents in order to reduce user-perceived latency when loading a page from a remote site (Srivastava et al., 2000). These rules are used in order to reveal correlations between pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. A transaction is a projection of a portion of the access log. In the work of Liu et al. (1998) and Wang et al. (2000) considered using association rules for prediction without consider sequential data. In the contrast, Lan et al. (1999) develops a specialized association rules mining algorithm to discover the prefetched documents. The counting of support is done differently than in Agrawal and Srikant (1994) since the ordering of documents is considered (Nanolopoulos et al., 2003). Only consecutive subsequences inside a user transaction are supported. For instance, the user transaction ABCD supports the subsequences: AB, BC, and CD. Yang et al. (2004) studied different association-rule based methods for web request prediction. In their analysis is based on a two dimensional picture and using real web logs as training and testing data. In this experiment found that the latest-substring produces the best prediction precision among all method and the pessimistic selection method is always the winner among the three rule-selection methods. In this work, we also used the latest-substring to represent the prediction rules.

2.3 Markov model

Markov models (Papoulis, 1991) have been used for studying and understanding stochastic processes, and were shown to be well-suited for modeling and predicting a user's browsing behavior on a web-site. In general, the input for these problems is the sequence of web-pages that were accessed by a user and the goal is to build Markov models that can be used to model and predict the web-page that the user will most likely access next. Padbanabham and Mogul (1996) use N-hop Markov models for improving pre-fetching strategies for web caches. Pitkow and Pirolli (1999) proposed a longest subsequence models as an alternative to the Markov model. Sarukkai (2000) use Markov models for predicting the next page accessed by the user. Cadez et al (2000) use Markov models for classifying browsing sessions into different categories. Markov model have been widely used to model user navigation on the web and predicting the action a user will take next given the sequence of actions he or she has already performed. For this type of problems, Markov models are represented by three parameters $< A, S, T >$, where A is the set of all possible *actions* that can be performed by the user; S is the set of all possible states for which the Markov model is built; and T is a $|S| \times |A|$ Transition Probability Matrix (TPM), where each entry t_{ij} corresponds to the probability of performing the action j when the process is in state i (Bestavros, 1996).

A Markov chain is a discrete-state random process in which the only state that influences the next state is the current state. To be more precise:

- Discrete-time Markov chain:

X_{n+1} depends only on X_n and not on any X_i, $1 \le i < n$

$$\Pr[X_{n+1} = S_i \mid X_n = S_j, X_{n-1} = S_k, ..., X_1 = S_l] = \Pr[X_{n+1} = S_i \mid X_n = S_j] \quad (1)$$

This equation is referred to as the Markov property.
- Continuous-time Markov chain:

Consider a continuous-time random process in which the number of times the random variables X(t) change value (the process changes state) is finite or countable. Let $t_1, t_2, t_3, \ldots$ be the times at which the process changes state. If we ignore how long the random process remains in a given state, we can view the sequence $\{X_{t_1}, X_{t_2}, X_{t_3}, ...\}$ as a discrete-time process embedded in the continuous-time process.

A continuous-time Markov chain is a continuous-time, discrete-state random process such that

(1) The embedded discrete-time process is a discrete-time Markov chain, and
(2) The time between state changes is a random variable with a memory-less distribution.

We describe a Markov chain as follows: We have a set of *states,* $S = \{s_1, s_2, \ldots, s_n\}$. The process starts in one of theses and moves successively from one state to another. Each move is called a *step*. If the chain is currently in state s_i, then it moves to state s_j at the next step with a probability denoted by p_{ij}, and this probability does not depend upon which states the chain was in before the current state.

The probabilities p_{ij} are called *transition probabilities*. The process can remain in the state it is in, and this occurs with probability p_{ii}. An initial probability distribution, defined on *S,* specifies the starting state. Usually this is done by specifying a particular state as the starting state.

3. Experimental design

In this paper, we study prediction models that prediction the user's next requests and compare prediction models by using association rules and Markov models for constructing the best prediction model. The prediction models that we build are based on web logs data that correspond with users' behavior. Therefore, they are used to make prediction for a general user and are not based on the data for a particular client. This requires the discovery of a web users' sequential access patterns and using these patterns to make predictions of users' future access. We will then incorporate these predictions into the web prefetching system in an attempt to enhance performance.

The experiment on the real data set is conducted to evaluate the performance of proposed framework. We used the web data, which collected from www.dusit.ac.th web server (figure 1) during 1 December 2004 – 31 December 2004. The total number of web pages with unique URLs is equal to 314 URLs. The web log records collected within 13,062 records are used to construct the user access sequences (figure 2). Each user session is split into training dataset and testing dataset. The training dataset is mined in order to extract rules, while the testing dataset is considered in order to evaluate the predictions made based on these rules. We experimentally evaluated the performance of the proposed approach: first-order markov model, second-order markov model, and association rule mining and construct the predictive model.

```
1102801060.863  1897600  172.16.1.98  TCP_IMS_HIT/304 203  GET http://asclub.net/images/main_r4_c11.jpg -
        NONE/- image/jpeg
1102801060.863  1933449  172.16.1.183  TCP_MISS/404 526  GET http://apl1.sci.kmitl.ac.th/robots.txt -
        DIRECT/161.246.13.86 text/html
1102801060.863  1933449   172.16.1.183  TCP_REFRESH_HIT/200 3565  GET
        http://apl1.sci.kmitl.ac.th/wichitweb/spibigled/spibigled.html - DIRECT/161.246.13.86 text/html
```

Figure 1: Web log data

3.1 Web log preprocessing

Web log files contain a large amount of erroneous, misleading, and incomplete information. This step is to filter out irrelevant data and noisy log entries. Elimination of the items deemed irrelevant by checking the suffix of the URL name such as gif, jpeg, GIF, JPEG, jpg, JPG. Because every time a Web browser downloads an HTML document on the Internet, several log entries since graphics and script are downloaded in HTML files too. In general, a user does not explicitly request all of the graphics that are in the web page, they are automatically down-loaded due to the HTML tags. Web usage mining is interesting to study the user's behavior, it does not make sense to include file requests that the user did not explicitly request. The HTTP status code returned in unsuccessful requests because there may be bad links, missing or temporality inaccessible pages, or unauthorized request etc: 3xx, 4xx, and 5xx. Executions of CGI script, Applet, and other script codes. Because it is not enough Meta data to map these requests into semantically meaningful actions, these records are often too dynamic and insufficiently information to make sense to decision makers.

3.2 Session identification

After removal, the log data are partitioned into user sessions based on IP and duration. The most of users visit the web site more than once time. The goal of session identification is to divide the page accesses of each user into individual sessions. The individual pages are grouped into semantically similar groups. A user session is defined as a relatively independent sequence of web requests accessed by the same user (Cooley et al., 2000). Fu et al. (1999) identify session by define the threshold idle time, if a user stays inactive for period longer than *max_idle_time*, subsequent page requests are considered to be in another episode, thus another session. Most of researcher use heuristic methods to identify the Web access sessions (Pallis et al., 2005) based on IP address and time-out does not exceed 30 minutes for the same IP Address, a new session is created when a new IP address is encountered or timeout. Catledge and Pitkow (1995) established a timeout of 25.5 minutes based on empirical data. In this research, we use IP address and time-out 30 minutes, if more than 30 minute then generate to new user (figure 2).

Session 1 : 900, 586, 594, 618
Session 2 : 900, 868, 586
Session 3 : 868, 586, 594, 618
Session 4 : 594, 618, 619
Session 5 : 868. 586. 618. 900

Figure2. User session from data set.

Assuming the access pattern of a certain type of user can be characterized by length of user transaction, and that the corresponding future access path is not only related to the last accessed URL. Therefore, users with relatively short transactions (e.g. 2-3 accesses per transaction) should be handled in a different way from users with long transactions (e.g. 10-15 accesses per transaction) (Wong et al., 2001). In this study, we propose a case definition design based on the transaction length. User transactions with lengths of less than 3 are removed because it too short to provide sufficient information for access path prediction (Wong et al., 2001).

3.3 Prediction using Association rules

The web log data is a sequence of entries recording which document was requested by a user. We extracted a prediction model based on the occurrence frequency and find the last-substring (Yang et al., 2004) of the W1. The last- substrings are in fact the suffix of string in W1 window. These rules not only take into account the order and adjacency information, but also the newness information about the LHS string. We used only the substring ending in the current time (which corresponds to the end of window W1) qualifies to be the LHS of a rule (Yang et al., 2004). For example, Table 1 shows the latest-substring rules.

Table 1. The latest-substring rules

W1	W2	The latest-substring rule
900, 586, 594	618	{594} ➔ 618
868, 586, 594	618	

From these rules, we extract sequential association rules of the form LHS□RHS from the session (Yang et al., 2004). The support and confidence are defined as follows:

$$supp = \frac{count\ (LHS \rightarrow RHS)}{number\ of\ sessions} \qquad (2)$$

$$conf = \frac{count\ (LHS \rightarrow RHS)}{count\ (LHS)} \qquad (3)$$

In the equation above, the function *count(Table)* returns the number of records in the log table, and *count(LHS)* returns the number of records that match the left-hand-side LHS of a rule.

3.4 Markov prediction model

The markov model has achieved considerable success in the web prefetching field (Pitkow and Pirolli, 1999; Deshpande and Karypis, 2001; Mobasher et al., 2002). However the limit of this approach in web prefetching is that only requested pages are considered. The state-space of the Markov model depends on the number of previous actions used in predicting the next action. The simplest Markov model predicts the next action by only looking at the last action performed by the user. In this model, also known as the first-order Markov model, each action that can be performed by a user corresponds to a state in the model (table 2). A somewhat more complicated model computes the predictions by looking at the last two actions performed by the user. It is called the second-order Markov model, and its states correspond to all possible pairs of actions that can be performed in sequence (table 3). This approach is generalized to the K^{th}-order Markov model, which computes the predictions by looking at the last K actions performed by the user, leading to a state-space that contains all possible sequences of K actions (Bestavros, 1996)

Table 2: Sample First-order Markov

1 st order	Support count					
Second item in sequence / First item in sequence	586	594	618	619	868	900
586	0	2	1	0	0	0
594	0	0	3	0	0	0
618	0	0	0	0	1	1
619	0	0	0	0	0	0
868	3	0	0	0	0	0
900	1	0	0	0	0	1

Table 3: Sample Second-order Markov

2 st order	Support count					
Second item in sequence / First item in sequence	586	594	618	619	868	900
586→594	0	0	2	0	0	1
594→619	0	0	0	1	0	0
868→586	0	1	1	0	0	0
900→868	1	1	0	0	0	0

Table 4: Prediction rule and confidence

Rule-selected	Prediction	confidence
586	594	2/3 = 67%
594	618	3/3 = 100%
868	586	3/3 = 100%
586→594	618	2/3 = 67%

3.5 Rule –selection

In our goal is to output the best prediction on a class based on a given training set. In rule-representation methods, each session where the previous visited matches the case can give rise to more than one rule. Therefore, we need a way to select among all rules that apply. In a certain way, the rule-selection method compresses the rule set; if a rule is never applied, then it is removed from the rule set. The end result is that we will have a smaller rule set with higher quality. In this section, we will study a method for rule selection. In addition to the extracted rules, we also define a default rule, whose the predict page is the most popular page in the training web log and the previous is the empty set. When no other rules apply, the default rule is automatically applied. For a given set of rules and a given rule-selection method, the above rule set defines a classifier. With the classifier, we can make a prediction for any given case. For a test case that consists of a sequence of web page visits, the prediction for the next page visit is correct if the predict of the selected rule occurs in prediction set. In association rule mining area, a major method to construct a classifier from a collection of association rules is the most-confident selection method (Liu et al., 1998). The most confident selection method always chooses a rule with the highest confidence among all the applicable association rules, among all rules whose support values are above the minimum support threshold. For example, suppose that for a testing set and a previous sequence is (A, B, C). Using the most-confident rule selection method, we can find 3 rules which can be applied to this example, including:

Rule 1: (A, B, C) ➔ D with confidence 35%
Rule 2: (B, C ➔ E with confidence 60%
Rule 3: (C) ➔ F with confidence 50%

In this case, the confidence values of rule 1, rule 2 and rule 3 are 35%, 60% and 50%, respectively. Since Rule 2 has the highest confidence, the most-confident selection method will choose Rule 2, and predict E.

4. Experimental results & discussions

The most commonly used evaluation metrics are accuracy, precision, recall, and F-Score to evaluate their classifier. Deshpande and Karypis (2004) used several measures to compare different Markov model-based techniques for solving the next-symbol prediction problem: accuracy, number of states, coverage and model-accuracy. In Haruechaiyasak (2003) and Zhu et al. (2002) used precision and recall to evaluate the performance of method. The precision measure the accuracy of the predictive rule set when applied to the testing data set. The recall measures the coverage or the number of rules from the predictive rule set that match the incoming request (Haruechaiyasak, 2003). To evaluate classifiers used in this work, we apply precision and recall, which are calculated to understand the performance of the classification algorithm on the minority class. Based on the confusion matrix computed from the test results, several common performance metrics can be defined as in Table, where TN is the number of true negative samples; FP is false positive samples; FN is false negative samples; FP is true positive samples. Precision and recall can be defined in term of:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

- *a* is the number of **correct** predictions that an instance is **negative**,
- *b* is the number of **incorrect** predictions that an instance is **positive**,
- *c* is the number of **incorrect** of predictions that an instance **negative**, and
- *d* is the number of **correct** predictions that an instance is **positive**.

Table 5: Confusion matrix

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

Several standard terms have been defined for the 2 class matrix:

- The *accuracy (AC)* is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a + b}{a + b + c + d} \tag{6}$$

- The *recall* or *true positive rate (TP)* is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = \frac{d}{c + d} \tag{7}$$

- The *false positive rate (FP)* is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$TN = \frac{a}{a + b} \tag{8}$$

- The *false negative rate (FN)* is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$FN = \frac{c}{c + d} \tag{9}$$

- Finally, *precision (P)* is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P = \frac{d}{b + d} \tag{10}$$

The accuracy determined using equation 6 may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases (Kubat et al., 1998).

4.1 Results

The results are plotted in figure 3 and show comparison of different algorithms. The experiments were conducted on dataset from www.dusit.ac.th web server. We evaluated our models using similar method to cross validation. We divided the web log as training data and testing data. As can be seen from the figure 3, the first-order Markov model gives the best prediction performance.

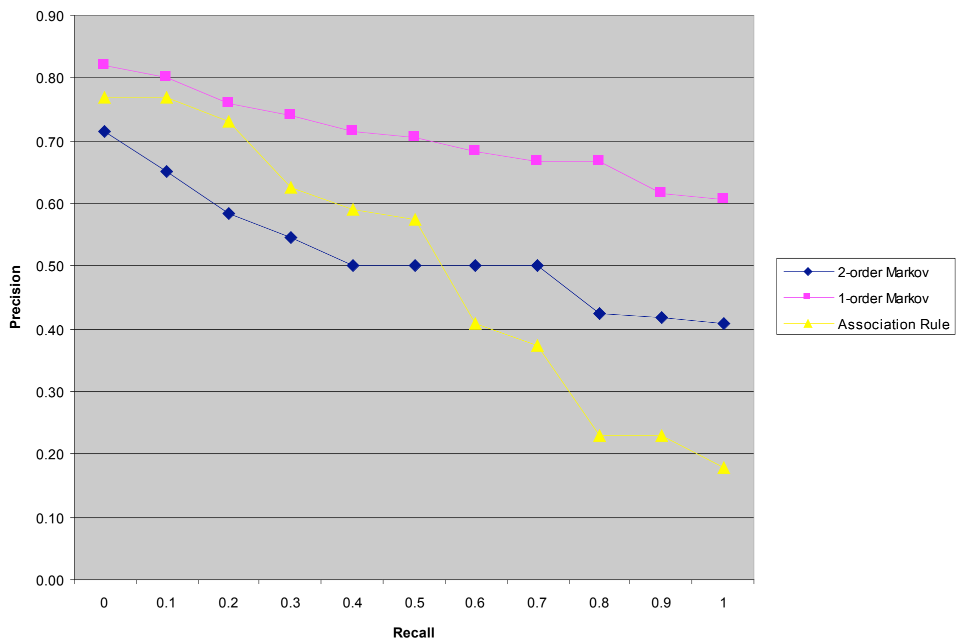


Figure 3. Result compare among three techniques

4.2 Discussions

Web usage mining is the application of data mining techniques to usages logs of large Web data repositories in order to produce results that can be used in the design tasks. In this experiment, the three algorithms are not successful in correct predicting the next request to be generated. This is because; in these models do not look far into the past to correctly discriminate the difference modes of the generative process.

5. Conclusions and future work

Today a large percentage of the Internet population accesses the World Wide Web. Users spend a lot of time impatiently waiting for the web pages to come up on screen. Web caching and Web prefetching are two important techniques used to reduce the noticeable response time perceived by users. Web prefetching technique utilizes the spatial locality of Web objects. In client's browser caching, Web objects are cached in the client's local disk. If the user accesses the same object more than one in a short time, the browser can fetch the object directly from the local disk, eliminating the repeated network latency.

Web servers keep track of web users' browsing behavior in web logs. Because of the increasing of users, World Wide Web is now suffering from the heavy traffic and the long latency. One way to reduce web traffic and speed up web accesses is through the use of prefetching. From log file, one can build statistical models that predict the users' next requests based on their current behavior. In this paper we studied different algorithm for web request prediction. We used proxy log that includes several kinds of client and contains a lot of client request from academic web servers. Our analysis based on three algorithm and using real web logs as training and testing data. Our conclusion is that the first-order Markov model is the best prediction among them. The Markov model construct of such an access model identifies the most likely hyperlink Web pages from the currently requested Web page. The first-order Markov model predicts the next action by only looking at the last action performed by the user.

In the future, we plan to use another algorithm and use the different method to extract sequence rules. However, each approach to prefetching assumes different environment, we must compare several approaches in each environment.

Acknowledgements

This work was supported in part by grants from Suan Dusit Rajabhat University, Thailand. http://www.dusit.ac.th

References

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceeding of the 20th VLDB Conference ages*. Santiago. Chile, 487–499.

Bestavros, A. (1996). Speculative Data dissemination and service to reduce server load. Network Traffic and Service Time. *Proceeding IEEE Conference Data Eng (IEEE ICDE'96)*, February, 151-160.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2000). Visualization of navigation patterns on a web site using model based clustering. Technical Report MSR-TR-00-18. Microsoft Research.

Catledge, L. & Pitkow, J.E. (1995). *Characterizing browsing behaviors on the World Wide Web*. Computer networks and ISDN Systems, 27(6).

Chakoula O. (2000). *Predicting users' next access to the web server*. Master thesis. Department of Computer Science, University of British Columbia.

Cooley, R., Tan, P-N., & Srivastava, J. (2000). Discovery of interesting usage patterns from Web data. Lecture Notes in Computer Science, 1836, 163-182.

Deshpande, M. & Karypis, G. (2001). Selective Markov models for predicting web page accesses. *In Workshop on Web Mining at the First SIAM International Conference on Data Mining*.

Deshpande, M. (2004). *Prediction/Classification technique for sequence and graphs*. Doctoral dissertation, University of Minnesota.

Etzioni, O. (1996). *The World Wide Web: Quagmire or Gold Mine*. Communications of the ACM, 39(11), 65-68.

Fu, Y., Sandhu, K., and Shih, M.Y. (1999). Clustering of Web users based on access patterns. *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. San Diego.

Haruechaiyasak, C. (2003). *A Data mining and semantic web framework for building a web-based recommender system*". Doctoral dissertation, University of Miami.

Lan, B., Bressan, S., Ooi, B.C., & Tay, Y. (1999). Making web servers pushier. *Proceeding Workshop Web Usage Analysis and User Profiling (WEBKDD '99).*

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association mining. *Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining*, 80-86.

Madria, S.K., Bhowmick, S.S., Ng, W.K. & Lim, E. (1999). Research Issues in web data mining. *Proceeding First International Conference on Data Warehousing and Knowledge Discovery*. Italy, Florence, 303-312.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002). Using sequential and non-sequential patterns in predictive web usage mining tasks. *Second IEEE International Conference on Data Mining (ICDM'02)*, 669-672.

Nanolopoulos A., Katsaros, D., & Manolopoulos, Y. (2003). A Data mining algorithm for generalized web prefetching. *IEEE Transactions on Knowledge and Data Engineering*, 5(5), September-October.

Padmanabhan, N. & Mogul, J.C. (1996). *Using predictive prefetching to improve World Wide Web latency*. *ACM SIGCOMM Computer Communication Review*, 26(3), 22-36.

Pallis, G., Angelis, L., & Vakali, A. (2005). *Model-based cluster analysis for web users sessions. Foundations of Intelligent Systems, 15th International Symposium (ISMIS 2005)*, Saratoga Springs, NY, USA, May 25-28.

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes*, NY: McGraw Hill.

Pitkow, J. and Pirolli, P. (1999). Mining longest repeating subsequences to predict World Wide Web surfing. *In Second USENIX Symposium on Internet Technologies and Systems*, 139-150.

Sarukkai, R.R. (2000). Link prediction and path analysis using Markov chain. *Proceeding of the 9th international World Wide Web conference on Computer networks: The international journal of computer and telecommunications networking*. Amsterdam. Netherlands, 377-386.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2), 12-23.

Su, Z. Yang, Q., Lu, Y., & Zhang H. (2000). What next: A prediction system for web requests using *N*-gram sequence models. *Proceeding of the First International Conference on Web Information Systems and Engineering Conference*, Hong Kong, pp.200–207.

Wang, K., Zhou, S.Q., & He, Y. (2000). Growing decision trees on association rules. *Proceedings of the International Conference of Knowledge Discovery in Databases*, 265-269.

Wong, C., Shiu, S., & Pal, S. (2001). Mining fuzzy association rules for web access case adaptation. *Proceeding of the workshop Programme at the fourth International Conference on Case-Based Reasoning*, Harbor Center in Vancouver, British Columbia, Canada.

Yang, Q., Li, T., &Wang, K. (2004). Building association-rule based sequential classifiers for Web-Document Prediction. *Journal of Data Mining and Knowledge Discovery*, 8, 253-273.

Zhu, J., Hong, J., & Hughes, J.G. (2002). Using Markov chains for link prediction in adaptive web site". *Proceeding of Soft-Ware 2002: First International Conference on Computing in an Imperfect World.* Lecture Notes in Computer Science, Springer, Belfast, April, 60-73.

Data Auditing by Hierarchical Clustering

Chun Hung Cheng, The Chinese University of Hong Kong
Chon-Huat Goh, Rutgers University, USA
Anita Lee-Post, University of Kentucky, USA

Abstract

Despite integrity checking at data entry, erroneous records still find their way to a database. Errors in a database impose a serious threat to an organization that uses the database. Therefore, data auditing is needed to examine the database and to identify erroneous records. This paper presents a data auditing approach based on hierarchical clustering techniques. Unlike some existing approaches, the data auditing approach proposed in this paper does not require knowledge of data. In addition, since this approach significantly reduces the number of parameter values to be tested, it is more efficient than existing approaches.

Keywords

Clustering, Data errors, Data auditing

Introduction

Errors in a database impose a serious threat to an organization that uses the database for information needs. Data errors are costly. Missing, incorrect, or incomplete data is the source of high operational costs, loss of good will, poor customer satisfaction and trust, disruption of business functions. More importantly, decisions based on poor quality of data cause long-term financial and competitiveness issues (Umar et al., 1999; Redman, 1998; Redman, 1995). To prevent erroneous records from entering a database, a database management system always examines incoming records for errors (Date, 1986; Wiederhold, 1983). For instance, data type checking ensures that data conform to their data types. Data range checking requires that the values of data fall between a legitimate range. However, these techniques cannot detect all erroneous records. In fact, according to The Data Warehousing Institute in Seattle, data quality problems costed U.S. businesses more than $600 billion per year in 2003 (Lais, 2003). PriceWaterhouseCoopers in New York surveyed 599 companies in 2001 and concluded that poor data management costed global businesses more than $1.4 billion per year in correcting billing, accounting and inventory errors (Betts, 2001). A case study of a number of Fortune 100 service organizations in 1998 found that one percent to five percent of data fields were erred causing eight percent to twelve percent loss of revenue and 40% to

60% increase in expense (Redman, 1998). Therefore, data quality is too important an issue to be ignored.

To maintain the quality of data, data auditing approaches examine a database periodically. These approaches identify records in a database that may contain errors in them and notify a database administrator (DBA) of these records. The DBA may then examine the identified records and correct their errors, if any.

There are two kinds of data auditing approaches. Application dependent approaches use knowledge of data and, therefore, require different auditing procedures for different database applications. On the other hand, application independent approaches do not need any prior knowledge of data and provide one general auditing procedure for all applications. The flexibility and generalizability of application independent approaches make them more appealing than the application dependent approaches. Hence, application independent approaches to data auditing are the foci of this paper.

Currently, application independent approaches use clustering analysis techniques. In data auditing, clustering analysis gathers records into groups or clusters based on their field values. Similar records occupy the same group while dissimilar records do not coexist in the same group. Records, whose field values make them significantly different from all others, may not find themselves related to any other group members at all (see Figure 1). They are called outliers (Storer and Eastman, 1990).

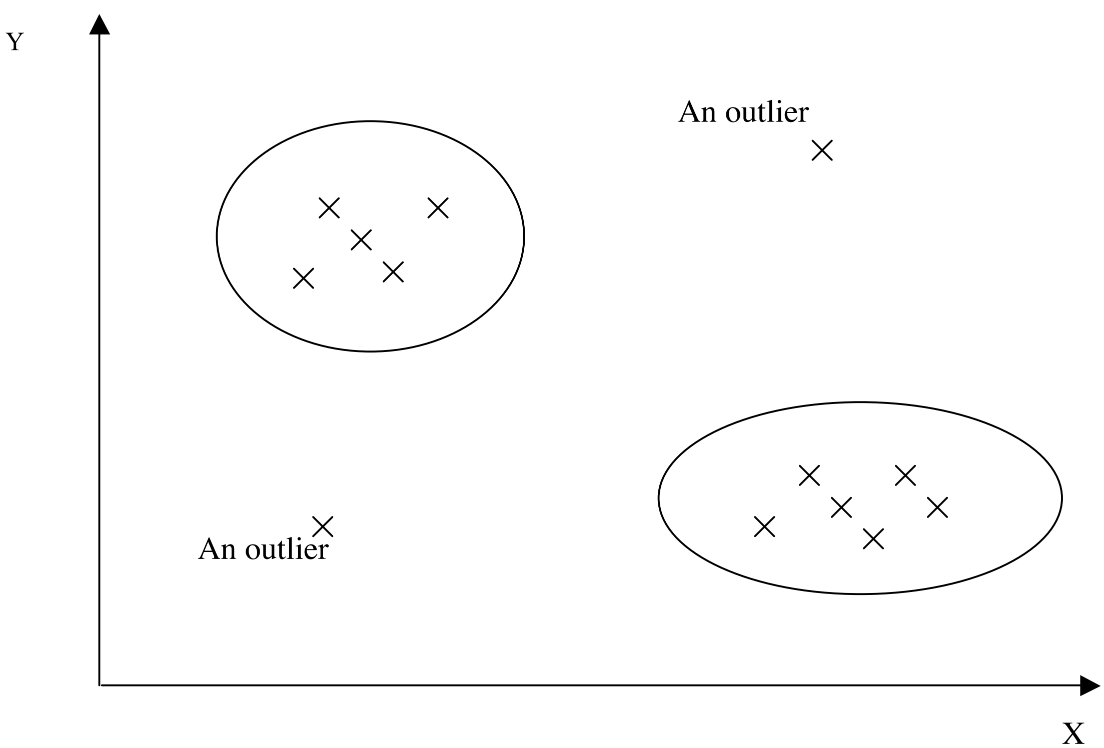


Figure 1: An Example of Two Groups and Two Outliers

Since records in the same group are similar, they are less likely to contain errors. On the other hand, an outlier contains irregular field value (or values) that may result from input, computational or other errors. In order to detect erroneous records, a good clustering approach must be able to identify such an outlier. A DBA then examines the outlier and determines its correctness.

In this paper, a hierarchical clustering technique is applied to data auditing. The new approach provides many advantages over current approaches.

1. Unlike some current approaches, the proposed approach does not require knowledge of data. Therefore, this approach is applicable to many different databases.
2. While many current approaches use non-hierarchical clustering techniques, the proposed approach uses hierarchical clustering techniques. All clustering approaches, including the proposed one, require the user to test the clustering algorithms with different parameter values. However, it will be shown in this paper that the proposed approach significantly cuts down on the search and evaluation of a number of parameter values and, therefore, is more efficient than other approaches.
3. Current approaches do not specify systematic ways to choose the best clustering result among many candidate solutions. The proposed approach applies a selection criterion to systematically evaluate the clustering of a database in order to best separate the outliers from similar records in a group. Therefore, the proposed approach provides a comprehensive implementable methodology to the data auditing problem.

The remainder of the paper is organized in the following manner. Section 2 discusses current approaches to data auditing and their limitations. In Section 3, a discussion on applying the hierarchical clustering technique to data auditing is provided with an illustrative example. Section 4 presents the conclusion.

Current Approaches of Data Auditing

Although the database literature considers errors in a database a serious problem (e.g., Fox et al., 1999; Umar et al., 1999; Felligi and Holt 1976; Naus et al. 1972), few studies propose ways to deal with the problem. There are two kinds of data auditing approaches: application dependent and application independent approaches.

Application Dependent Approaches

Application dependent approaches such as those by Schaffer (1987), Felligi and Holt (1976), Naus et al. (1972), and Freund and Hartley (1967), are all statistical-based. In detecting errors in a database, these approaches require knowledge of the data. Using these approaches, software developers may have to develop different programs for different database applications.

Application Independent Approaches

All application independent approaches use clustering analysis techniques. Lee et al. (1978) first applied a clustering approach to data auditing. They defined a distance function to measure the difference between two records. Based on a distance matrix, they found the shortest path between a pair of records. Since the determination of the shortest path is an NP-complete problem (Storer and Eastman, 1990), the shortest spanning path algorithm (Slagle et al., 1975) was used to find an approximate solution. A link between two records that is longer than the pre-specified threshold value will be broken. Records whose distances are less than the threshold value are similar and are placed in the same group. A record with no similar partners is an outlier.

Storer and Eastman (1990) proposed three related clustering approaches. They used the same distance function as defined by Lee et al. (1978). The first approach is called the leader algorithm (Hartigan, 1975). The leader algorithm clusters M records into K groups, where M and K are positive integer values and $M \geq K$. It assumes that the distance function between two records and the threshold value for group membership are available. The first record is a leader for the first group. A record is assigned to an existing group if its distance from the group leader is less than the threshold value. It becomes a new leader for a new group if its distance from every existing leader is more than the threshold value.

The second approach is a modification of the leader algorithm, which we refer to as an average record leader algorithm. This modified algorithm uses the average record instead of the first record as an initial leader. Therefore, the algorithm can generate a solution independent of record order. On each pass, a record that is furthest from its group leader becomes a leader for a new group. If the algorithm were to produce K groups, it requires K passes through the data.

The third approach is another modification of the leader algorithm. Storer and Eastman (1990) call it the greatest distance algorithm. The greatest distance algorithm uses a different criterion for selecting new group leaders. First, Storer and Eastman (1990) define a non-deviant cluster as one that has more than one percent of all records. A new leader is the record that is furthest from a leader of a non-deviant cluster and is greater than the average record distance from its cluster leader.

Limitations of Current Approaches

Table 1 summarizes the parameters that need to be pre-defined for all four approaches. The shortest spanning path algorithm requires pre-specifying a threshold value. Since there is no upper bound on the value of the distance function, there are many possible threshold values to test before a desirable value is found. Worse yet, there is no systematic way to find the most desirable value. The search for this parameter value may be time-consuming. A smaller threshold value than the desirable may result in a larger number of groups and possibly a larger number of outliers that are actually error free. A larger threshold value than the desirable, on the other hand, may lead to a fewer number of groups and group erroneous records into existing groups along with correct records.

Approaches	Pre-specified threshold value	Pre-specified number of groups
SSP	Yes	No
LA	Yes	Yes
ARLA	No	Yes
GDA	No	Yes

Notation:

SSP The shortest spanning path algorithm
LA The leader algorithm
ARLA The average record leader algorithm
GDA The greatest distance algorithm

Table 1. The required parameters of data auditing approaches

The leader algorithm also uses a pre-specified threshold value and the number of groups to be formed. It is difficult for the user to come up with the desirable values for both the threshold value and the number of groups to be formed as the number of possible combinations for each value is very large. The desirable values can be found only after an extensive search. The remaining two algorithms need the pre-specified number of groups. A larger number of groups than desirable produces many outliers containing no errors while a smaller number misses some outliers containing errors since we do not have prior knowledge of the data, we have to test the algorithms with many parameter values.

Neither Lee et al. (1978) nor Storer and Eastman (1990) discuss how to find the desirable parameter values among many possible values. Moreover, they do not specify any systematic methods to choose the desirable clustering result among many candidates that are generated from a given set of parameter values. These problems are specifically addressed in this paper.

Hierarchical Clustering

In this section, the proposed approach of hierarchical clustering to data auditing is presented with an illustrative example. This approach involves four major steps, which are as follows:

1) definition of distance functions,
2) formation of a dendogram using a hierarchical clustering technique,
3) generation of threshold values based on the results of the dendograms, and
4) identification of outliers based on a selection criterion.

Definition of distance functions

A record, R_i, in a database consists of more than one field. R_i can be represented in the form of a vector. That is, $R_i = (x_{i1}, x_{i2}, x_{i3},...,x_{iN})$, where x_{ip} is the value of the *pth* field of R_i, for p =1, 2, …. N and i = 1, 2, …, M. Based on the field values of a record, the record can be classified into one of the following three types: (Lee et al., 1978)

A. Type I records: All field values in this type of record are numerical. The distance between two records R_i and R_j is defined as

$$d_{ij} = \sum_{p=1}^{N} c(x_{ip}, x_{jp}) / N ,$$

where

$$c(x_{ip}, x_{jp}) = | x_{ip} - x_{jp} | / S_p \text{ , and}$$

$$S_p = | \max_{1 \le i \le M} x_{ip} - \min_{1 \le i \le M} x_{ip} | .$$

For example, if R_i = (4.5, 3.1, 0.9, -2.1), R_j = (4.1, 2.1, 0.3, -1.1), S_1 = 5.0, S_2 = 4.0, S_3 = 2.0, and S_4 = 2.1, then d_{ij} = 0.2765.

B. Type II records: All field values in this type of record are non-numerical. The distance between two records R_i and R_j is defined as:

$$d_{ij} = \sum_{p=1}^{N} c(x_{ip}, x_{jp}) / N$$

where

$$c(x_{ip}, x_{jp}) = \{ \begin{matrix} 1\ if\ x_{ip} \neq x_{jp} \\ 0\ otherwise. \end{matrix}$$

C. Type III records: Fields in a type III record may assume either numerical or non-numerical values. The distance between two records R_i and R_j is defined as:

$$d_{ij} = \sum_{p=1}^{N} c(x_{ip}, x_{jp}) / N$$

where for a numerical field p,

$$c(x_{ip}, x_{jp}) = | x_{ip} - x_{jp} | / S_p \text{, and}$$

$$S_p = | \max_{1 \le i \le M} x_{ip} - \min_{1 \le i \le M} x_{ip} |$$

or for a non-numerical field p,

$$c(x_{ip}, x_{jp}) = \{ \begin{matrix} 1\ if\ x_{ip} \neq x_{jp} \\ 0\ otherwise. \end{matrix}$$

For example, if R_i = (black, black, 3.1, 5.0), R_j = (black, white, 2.1, 5.1), S_3 = 4.0, and S_4 = 5.5, then d_{ij} = 0.3170.

Lee et al. (1978), and Storer and Eastman (1990) use Euclidean distances or city block distances for type I records, and hamming distances for type II records. There is no upper bound on the value of either distance function. Therefore there are a large number of possible threshold values. The proposed approach uses a new distance function that assumes a value between 0 and 1 inclusively, thus significantly reducing the number of possible threshold values. In addition, a hierarchical clustering technique for clustering is used here to further reduce the number of possible threshold values.

To illustrate the new distance function, consider a simple example with a type III record. Table 2 is a personnel database for a hypothetical company. Fields POSITION, and EDUCATION are non-numerical while fields MONTHS, and SALARY are numerical.

Record ID	POSITION[1]	EDUCATION[2]	MONTHS[3]	SALARY[4]
1	0	0	15	20,000
2	1	1	10	20,000
3	0	0	11	20,000
4	1	1	35	60,000
5	1	0	17	30,000
6	0	1	17	30,000
7	0	0	16	20,000
8	1	1	33	65,000
9	1	0	16	46,000
10	0	0	50	80,000

Note
1. Field POSITION = 1, when an employee has a middle management position; and Field POSITION = 0, when an employee has a supervisor position.
2. Field EDUCATION = 1, when an employee has a college degree; and Field EDUCATION = 0, when an employee does not have a degree.
3. Field MONTHS is the number of months an employee has worked for the company.
4. Field SALARY is the current salary of an employee.

Table 2. Salary history of ten employees

Distance matrix (1) is constructed by evaluating the distance functions of all pairs of employee records. Only entries above the diagonal are shown because $d_{ii} = 0$ and $d_{ij} = d_{ji}$. An element in the distance matrix represents the extent to which a record is different from another one.

Records

Records	1	2	3	4	5	6	7	8	9	10
1		.52	.02	.73	.29	.29	.01	.73	.34	.36
2			.51	.25	.32	.32	.53	.26	.36	.89
3				.74	.31	.31	.03	.75	.36	.89
4					.43	.43	.72	.03	.39	.64
5						.50	.29	.44	.06	.57
6							.29	.44	.56	.57
7								.73	.33	.36
8									.39	.63
9										.53
10										

(1)

Hierarchical Clustering Using the Distance Function

A hierarchical clustering technique operates on a distance matrix. It constructs a dendogram that depicts relationships among records. Anderberg (1973) discusses seven hierarchical clustering techniques. Among the seven techniques, single linkage, average linkage, and complete linkage clustering are most widely used. In this paper, we choose single linkage clustering to illustrate the example given in Section 3.1. Note that other hierarchical clustering techniques may also apply.

The single linkage clustering algorithm operates on distance matrix (1) and produces the dendogram in Figure 2.

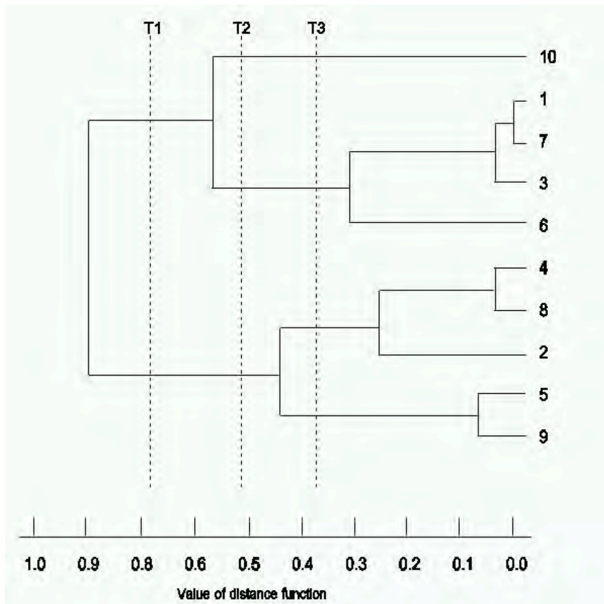


Figure 2. The dendogram of distance matrix (1)

Generation of Threshold Values Using Dendograms

At this stage, a threshold value is needed to place records into groups. With the help of the dendogram, we can significantly reduce the number of possible threshold values to be examined. For example, Figure 2 shows the three possible threshold values and they are indicated as T1, T2, T3. The highest and second highest values of distance function are 0.89 and 0.57, respectively. A threshold value such as T1, where 0.57 ≤ T1 ≤ 0.89, forms two groups with no outliers (i.e. {1,3,6,7,10}, and {2,4,5,8,9}). Similarly, a threshold value T2 where 0.57 ≤ T2 ≤ 0.44 finds three groups (i.e. {1,3,6,7}, {10}, {2,4,5,8,9}) with record 10 as an outlier.

Identification of outliers

All clustering results obtained by using various threshold values are given in Table 3. Since each result has its own associated outliers, a selection criterion is needed to determine the best clustering result so that the most appropriate outliers can be identified.

The above expression is the selection criterion used in determining the best clustering result. The value of this selection criterion for each clustering result of our example is given in Table 3. Among the eight clustering results, solution 5 is the best as the value of its selection criterion is the lowest. Based on solution 5, we conclude that records 2, 6, and 10 are outliers. A DBA will examine these outliers and determine whether they contain any error or not. Our example postulates that an employee with a college degree, in a middle management position, and with more years of seniority should have higher current salary. On the other hand, an employee without a college degree, in a supervisor position, and with fewer years of seniority should have a lower current salary. However, record 2 indicates that the employee has a college degree and in a middle management level but has relatively low current salary. Record 10 indicates that the employee has exceptionally high current salary for his/her position (i..e., supervisor) and education (i.e., does not have a degree). Record 6 may (or may not) contain errors.

Solution No.	Clustering results	Outlier/s	Value of criterion
1	{1,3,6,7,10}, and {2,4,5,8,9}	no outlier	-0.2581
2	{1,3,6,7}, {2,4,5,8,9}, and {10}	10	-0.2775
3	{1,3,6,7}, {2,4,8}, {5,9}, and {10}	10	-0.3439
4	{1,3,7}, {2,4,8}, {5,9}, {6}, and {10}	6,10	-0.3897
5	{1,3,7}, {4,8}, {5,9}, {2}, {6}, and {10}	2,6,10	-0.4431
6	{1,3,7}, {4,8}, {2}, {5}, {6} {9}, and {10}	2,5,6,9,10	-0.4402
7	{1,3,7}, {2}, {4}, {5}, {6}, {8}, {9}, and {10}	2,4,5,6,8,9,10	-0.4322
8	{1,7}, {2}, {3}, {4}, {5}, {6}, {8}, {9}, and {10}	2,3,4,5,6,8,9,10	-0.4244

Table 3. The possible clustering results

Conclusion

In this paper, the limitations of current data auditing approaches are first discussed. To overcome these limitations, a hierarchical clustering approach to data auditing is proposed.

Unlike many existing statistical-based data auditing approaches, the proposed approach is application independent and does not require prior knowledge of data. Thus, the proposed approach can be used to detect erroneous records in any database application. In addition, it significantly cuts down on the search and evaluation of parameter values and, therefore, is more efficient than other approaches. Furthermore, it provides a systematic and consistent evaluation of clustering results. Taken together, the proposed hierarchical clustering approach is a comprehensive, flexible, and easily implementable methodology to the data auditing problem.

References

Anderberg, M.R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Betts, M. (2001). Dirty Data. *ComputerWorld*. December 17.

Date, C.J. (1986). *An introduction to database systems*. California: Addison-Wesley.

Felligi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Editing and Imputation. *Journal of American Statistics Association*, (71), 17-35.

Fox, T.I., Guynes, C.S., Prybutok, V.R., and Windso, J. (1999). Maintaining Quality in Information Systems. Journal of Computer Information Systems, Fall, 76-59.

Freund, R.J. and Hartley, H.O. (1967). A Procedure for Automatic Data Editing, *Journal of American Statistics Association*. (62) 1967, pp. 341-352.

Hartigan, J.A. (1975). *Clustering Algorithms*. New York: McGraw-Hill.

Lais, S. (2003). Piecing Together the Data Picture. *Computerworld*. August 11.

Lee, R.C., Slagle, R.R. and Mong, C.T. (1978). Towards Automatic Auditing of Records. *IEEE Transactions on Software Engineering* (4), 441-448.

Naus, J.I., Johnson, T.G., and Montalvo, R. (1972). A Probabilistic Model for Identifying Errors and Data Editing. *Journal of American Statistics Association*. (67), 943-950.

Redman, T.C. (1998). The Impact of Poor Data Quality on the Typical Enterprise. *Communications of the ACM*, (41), 79-82.

Redman, T.C. (1995). Improve Data Quality for Competitive Advantage. *Sloan Management Review*, (36), 99-107.

Schaffer, J. (1987). Procedure for Solving the Data-Editing Problem with Both Continuous and Discrete Data Types. *Naval Research Logistics*, (34), 879-890.

Slagle, J.R., Chang, C.L., and Heller, S.R. (1975). A clustering and data-reorganizing algorithm, *IEEE Transactions on Systems, Man, and Cybernetics*, (5), 125-128.

Stanfel, L.E. (1983). Applications of Clustering to Information System Design. *Information Processing & Management*, (19), 37-50.

Storer, W.F. and Eastman, C.M. (1990). Some Experiments in the use of clustering for data validation. *Information Systems*, (15), 537-542.

Umar, A., Karabatis, G., Ness, L., and Horowitz, B. (1999). Enterprise Data Quality: a Pragmatic Approach. *Information Systems Frontiers*, (1), 279-301.

Wiederhold, G. (1983). Database design. New York: McGraw-Hill.

Correspondence for the author may be sent to Dr. Anita Lee-Post, School of Management, Gatton College of Business and Economics, University of Kentucky, Lexington, KY 40506.
E-mail: dsianita@uky.edu

Machine assignment in cellular manufacturing layout using genetic algorithm

V. Anbumakar, Anna University, India
G. Prabhaharan, National Institute of Technology (formerly REC), India.

ABSTRACT

This paper presents the problem of machine assignment in cellular manufacturing layout. The focus is only on the intra-cell (within cell) machine assignment with the objective of minimizing the intra-cell movement by assigning the machines in the optimum position within a 3x3 grid. Genetic Algorithm (GA) is proposed to solve the problem. To examine the performance of the proposed GA, data sets taken from the literature are used and the results are compared with that of from other approaches. The proposed GA provides a near optimal layout.

Keywords

Genetic Algorithm, Machine assignment, Cellular manufacturing layout.

1. Introduction

Cellular manufacturing (CM) is an established international practice to integrate: equipment, people and systems into "focused -factories", "mini-businesses", or "cells" with clear customers, responsibilities and boundaries.

CM is implemented in three stages:

1. **Cell formation** – Grouping the parts in to part families and machines in to cells by the parts production process.
2. **Facility layout** – Layout of cells within the shop floor (inter-cell layout) and layout of machines within each cells (intra-cell layout)
3. **Scheduling** - Scheduling of jobs in each cells.

The major elements in exploiting the benefits of CM by efficient layout design are

- Setup time reduction
- Work-in-progress inventory reduction
- Material handling cost reduction
- Reduction in material flow distance
- Improvement in machine utilization
- Improvement in quality

The facility layout in cellular manufacturing systems involve the arrangement of cells within the floor space, so as to minimize the inter-cell layout movement. The machine layout in cellular manufacturing systems involves the arrangement of machines within the cells so as to minimize the intra-cell movement. Three basic types of machine layout is identified by Hassan.M.M.D (1994) are

1. **Single row layout**: In this layout, different types of machines may be arranged in a single row as close as possible to the sequence of operations. The layout may be assuming several shapes such as linear, semi-circular or U-shaped.
2. **Multi-row layout**: The machines are arranged in more than one row in this type of layout. The machines in each row interacts with each other and as well as with the machines in other rows.
3. **Loop layout**: In this layout, the machines are arranged around an oval path and the movement of parts is usually unidirectional

The layout problem in manufacturing system was considered by Tam.K.Y. (1991) and presented a hierarchical approach, which consist of three phases: (i) cluster analysis (ii) initial layout (iii) layout refinement. Yaman.R (1993) described an effective sorting method, which provided a direct method to construct a layout for further improvement and measured the effectiveness of the method with case studies. Meller.R.D (1996) presented recent and emerging trends in the facility layout problem, including new methodology, objectives, algorithms and extensions to the problems. Welgama.P.S. (1996) proposed an integrated methodology using Knowledge-based Optimization approach for automating the determination of layout and material handling system.

Later Dweiri.F. (1996) proposed a methodology, based on fuzzy set theory, to improve the facility process. Chiang.W.C. (1996) presented an improved Tabu search heuristic for facility layout design problem. Chiang.W.C. (1998) implemented a set of heuristic procedures like Tabu search, a probabilistic Tabu search heuristic, a SA heuristic and a hybrid Tabu search heuristic to solve facility layout problems. Castillo.I (2003) proposed a heuristic algorithm to solve an extended distance-based facility problem that concurrently determines the number and shape of the department, the assignment of machines to the departments and allocation of part flow volume to individual machines.

The facility layout problem in CM has not captured researcher's attention as much as cell formation. The layout problem however plays an important role in the design of CM. Only a few researchers have dealt with this subject. Jajodia.S (1992) presented a new method CLASS (Computerized Layout Solutions using Simulated annealing) that considered the inter-cell layout and intra-cell layout problem in cellular manufacturing environment and addressed the relative placement of equidimensional manufacturing entities with in a discrete solution space in an attempt to minimize the total material flow between these entities. Souilah.A (1995) and Wang.T.Y (1998) proposed the Simulated Annealing algorithm to solve facility layout problems in cellular manufacturing system. Hassan.M.M.D. (1995) reviewed and consolidated the emerging literature on the Group technology layout. He suggested a framework of analysis for developing the Group technology layout. Bazargan-Lari.M (1999, 2000) presented the application of recently developed design methodologies in cellular manufacturing environment.

The machine layout has received substantially less attention even though it is related to facility layout problem. Heragu.S.S (1990) addressed the machine layout problem in automated manufacturing system and presented a Knowledge Based system for Machine Layout (KBML), which is a combination of optimization and expert system approach to solve the problem. Hassan.M.M.D (1994) reviewed the machine layout problems in modern manufacturing facilities and addressed the design issues of machine layout. Also suggested some optimization approaches to develop the layout. Tanchoco.J.M.A. (1999) introduced a facility layout design procedure to find the best machine grouping along with location of pickup and delivery stations and machine layout for each cell based on Segmented Flow Topology (SFT) with the objective of minimizing the total material handling cost. Urban.T.L (2000) proposed a model that does not require the machines to be placed in a functional layout or in a cellular manufacturing arrangement but allows the material flow requirements to dedicate the machine placement. W.M.Chan (2002) presented a heuristic algorithm to solve the problem of machine allocation in cellular manufacturing with the objective of minimizing the total traveling score. This algorithm used an adaptive approach to relate the machine in a cell by examining the merged part flow weights of machine pair.

In this work the implementation of Genetic Algorithm (GA) is attempted to solve the machine assignment problem, with the objective that minimizes the total traveling distance.

2. Formation of the Model

This research aims in developing a model to determine the total distance traveled by parts, when the machines are assigned in a cellular manufacturing layout. The machine location site is a regular grid; similar to W.M.Chan (2002)], each machine cluster contains machines requiring equal floor space areas. The maximum number of machines assigned to a grid is limited to 9 machines as a 3x3 grid and there is only 9 potential machine location zones as shown in figure.1

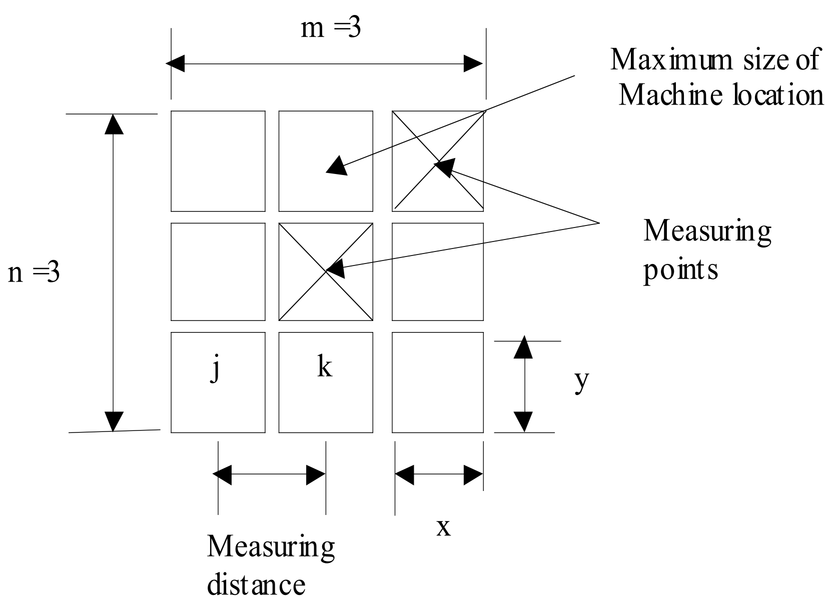


Figure 1: Nine zones location site

This paper focuses on applying Genetic algorithm to solve the machine assignment problem in cellular manufacturing systems. The assumptions are to be made for solving the problem:

1. The cell formation is completed first, i.e. the machines belongingness to cell are known
2. Size of any machine location zone (x, y) can accommodate the largest machine in that particular cell.
3. A machine can be assigned to any location.
4. The loading and unloading point is at the center of each machine.
5. The distance between any two neighboring machines in the cell is equal.
6. Traveling distance is estimated by measuring the rectilinear distance from center of the machine to the center of the destination machine.

The main objective of this model is to minimize the total traveling distance in a period and it is defined as

$$\text{Minimize } \zeta_p = \sum_{j=1}^{m} \sum_{k=j+1}^{m} D_{j \leftrightarrow k} * Q_{p=i}$$

Where,

$$D_{j \leftrightarrow k} = |X_j - X_k| + |Y_j - Y_k|$$

P: Planning period, P= 1… P;
i: Type of part, i = 1,….n;
j ,k : a machine pair, j,k =1,….m; j ≠ k
Q: Quantity demand of part in a planning period;
D: Distance between a machine pair;
ζ : Total traveling distance within a cell in a period;

3. GENETIC ALGORITHM

Genetic Algorithm (GA) proposed by J.H.Holland (Kalyanmoy Deb 2002) is a stochastic search method that mimics the metaphor of natural biological evolution. GA operates on a population of potential solutions applying the principles of survival of the fittest to produce better and better approximation solutions. At each generation, a new set of approximation is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators borrowed from natural genetics. This process leads to the evolution of population of individuals that are better suited to their environment than the individuals that they were created from, just as in natural adaptation. The fundamental underlying mechanism of GA operates upon a generation and generally consists of three main operators.

1. **Reproduction**: Selection of copies of chromosome according to the fitness value.

2. **Crossover**: Exchange of portions between chromosomes.
3. **Mutation**: A random modification of chromosomes.

3.1. PROPOSED GA FOR MACHINE ASSIGNMENT

GA based optimization model for machine assignment in cellular manufacturing system is shown in the flow chart in figure 2 and detailed procedures are explained as follows.

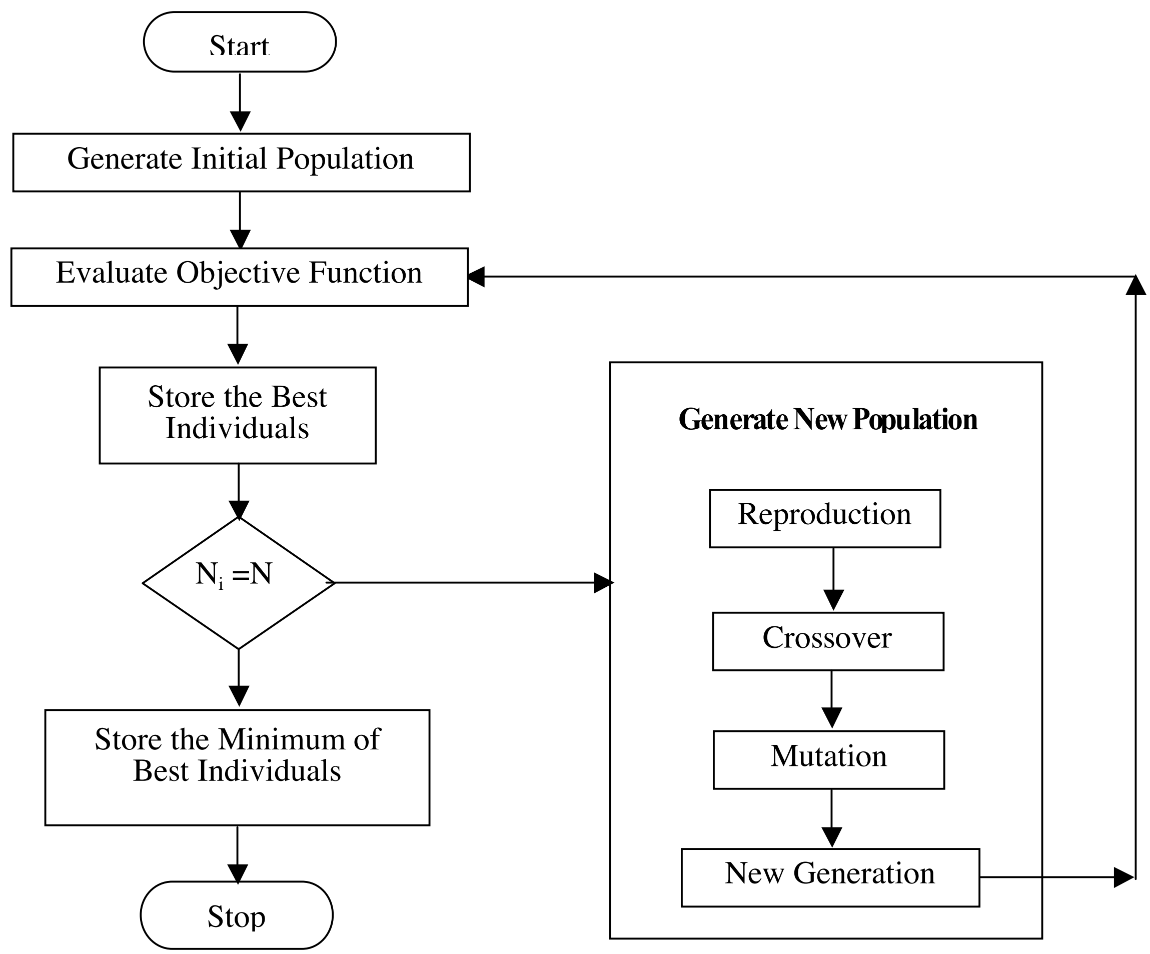


Fig. 2 Flow Chart of proposed Genetic Algorithm

Step: 1 Representation:

The length of a chromosome represents the number of machines considered in the problem. The gene in the chromosome indicates the machine number and the position number indicates the machine present in that position. The position of any machine represents its location.

Example: 2 4 1 3 5 8 9 7 6

Here the length of the chromosome is 9, which indicates that the number of machines considered in the problem is 9. The genes in the chromosome take the values 1 to 9 which indicates the machine number. Position of machine 2 is 1, 4 is 2 and so on.

Step: 2 Initialization:

Randomly generated population of 40 chromosomes is used as initial population. Each chromosome is converted in to a 3x3 matrix and the objective value is calculated using the objective function equation.

Example: 241358976

$$\begin{bmatrix} 241 \\ 358 \\ 976 \end{bmatrix}$$

Step: 3 Fitness function calculation:

Since the objective of this problem is minimization the following fitness equation is used to calculate the fitness function.

$$\text{F(x)} = \frac{1}{1 + f(x)}$$

Where f (x) is the objective function value of each chromosome.

Step: 4 Selection and reproduction:

In this process, the probability of selecting the string from initial population is using the following relation

$$\text{P}_i = \frac{f(i)}{\sum_{j=1}^{n} f(i)}$$

and the cumulative probability is also found out.
The Roulette wheel selection is used for reproduction by generating random numbers.

Step: 5 Cross over:

The single point cross over method is used for cross over process with a cross over probability of 0.6. The cross over point is randomly selected. After the cross over process, the machine number is checked. If any machine number is repeated, replace that machine numbers with another machine number, which is not in the chromosome.

Example:

Parent 1: 241358976

Parent 2: 134579862

Cross over point: 4 (Randomly selected)

After cross over:

Child 1: 241379862

Child 2: 134558976

After cross over machine number 2 is repeated in child 1. So machine number 2 is replaced with 5, which is not in the chromosome. Similarly in child 2 machine number 5 is replaced with 2.

Final chromosomes:
1: 241379865
2: 134258976

Step: 6 Mutations:

Mutation is done with a mutation probability of 0.04 using Swapping operation. The mutation location is selected randomly.

Example: 2 4 1 3 5 8 9 7 6
Mutation location: 3, 7 (Randomly selected)

After mutation: 2 4 9 3 5 8 1 7 6

Step: 7 Start the next generations

Step: 8 Terminate after the given number of generations are carried out.

4. NUMERICAL ILLUSTRATION

Here, the problem by W.M.Chan (2002) is considered, to which GA is applied. The problem involves 5 parts that are processed in 9machines.Details about the operational sequence and quantitative demands for a 5 period planning horizon are furnished in table 1(a) and table 1(b).

Types of Parts	Operation sequence of parts
1	1-3-5-7-2-7-9
2	1-4-2-5-6-8-9
3	1-5-7-8-5-6-2-9
4	1-2-4-6-7-8-2-3-9
5	1-7-6-4-2-8-3-5-6-9

Table 1 (a) Operational Sequences

Period Types of Parts	**Period 1**	**Period 2**	**Period 3**	**Period 4**	**Period 5**
1	10	35	90	40	55
2	30	50	25	65	20
3	45	15	40	70	15
4	70	80	55	90	85
5	85	60	70	20	30

Table 1(b) Quantitative Demand

5. Results and discussion

For evaluation purposes, all computations are programmed in the C^{++} language and run on a PC with Pentium IV CPU. To confirm the solution quality of proposed GA, the test problem taken from the literature is solved. The results of the problems are shown in figure 3, which shows the arrangement of machines with in 3x3 grid and total traveling distance for each period.

1	5	3	1	2	4	4	2	1	3	1	5	9	6	4
2	7	8	7	8	6	6	7	5	4	2	6	8	7	2
4	6	9	9	3	5	9	8	3	9	8	7	3	5	1
ζ = 2780			**ζ = 2680**			**ζ = 2950**			**ζ = 3125**			**ζ = 2200**		

Figure3: The arrangement of machines and total traveling distance for each period.

These results are compared with the optimal solution and the other approaches proposed by Yaman.R (1993), Tang.C (1996) and Chan.W.M. (2002) using the same data and the same 3X3 grid location are shown in figure 4 and table2. The results showed that the proposed GA performs within 1%variation from optimal solution and better than the other approaches. The convergence of GA is shown in figure 5, 6, 7, 8 & 9.

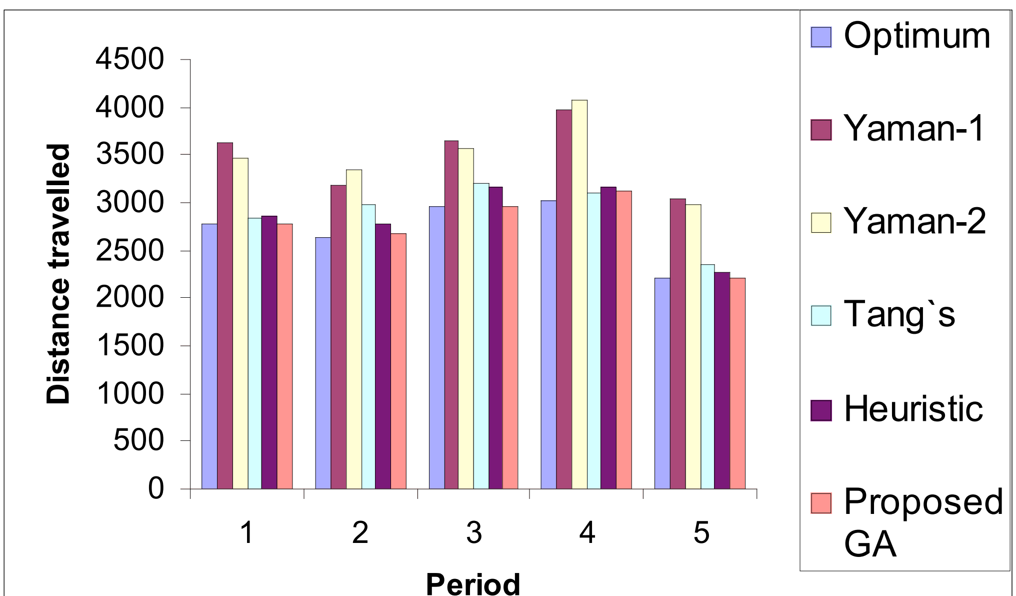


Figure 4: Comparison of results

Period	Optimum	Yaman's spiral 1	Yaman's spiral 2	Tang's approach	Heuristic algorithm	Proposed GA
1	2780	3630	3470	2820	2850	2780
2	2640	3180	3350	2980	2790	2680
3	2950	3650	3570	3200	3160	2950
4	3020	3975	4065	3100	3165	3125
5	2200	3045	2975	2355	2275	2200
Total	13,590	17,520	17,430	14,455	14,240	13,735
% of variation with optimal	-	28.92	28.26	6.36	4.78	1.05
% of variation with proposed GA	-	21.60	20.79	4.98	5.54	-

Table 2: comparison of results

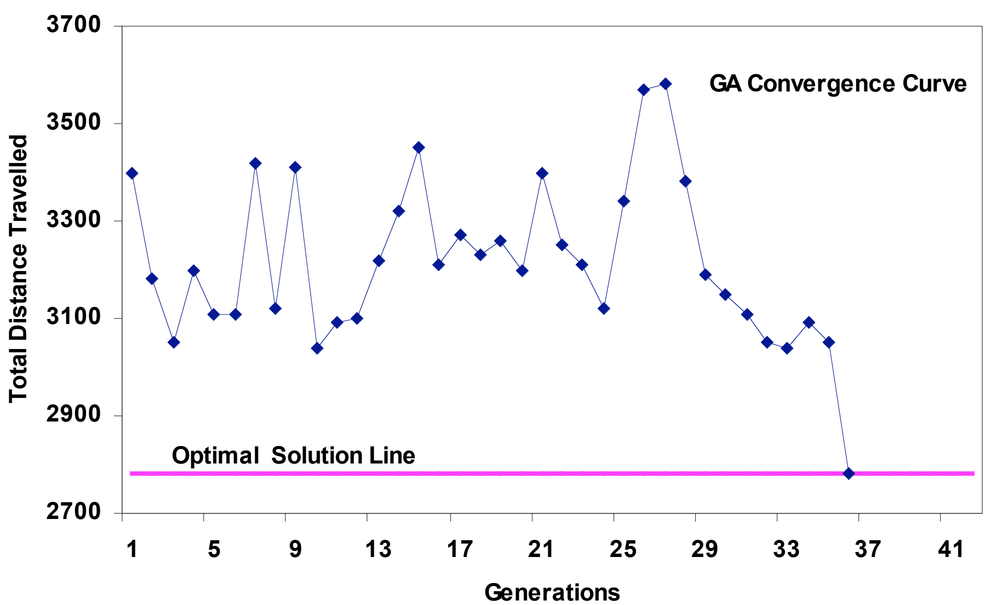


Figure 5:GA convergence for period 1

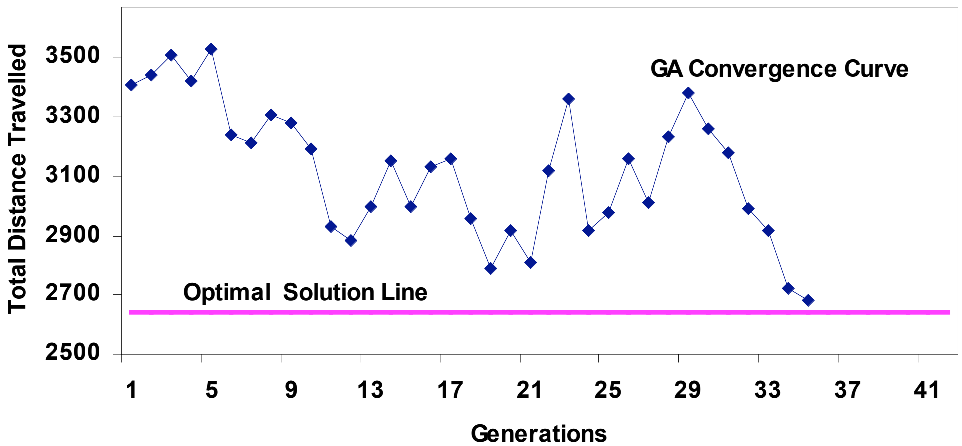


Figure6: GA convergence for period 2

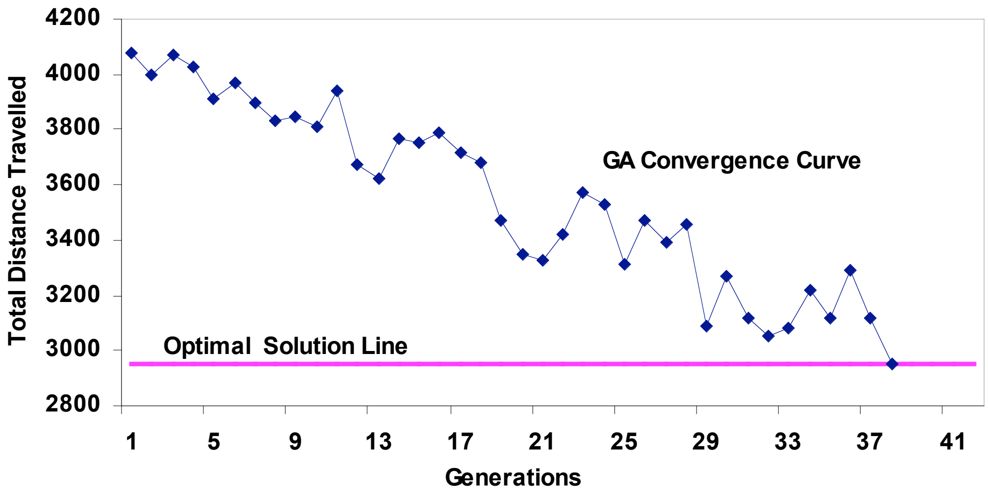


Figure7: GA convergence for period 3

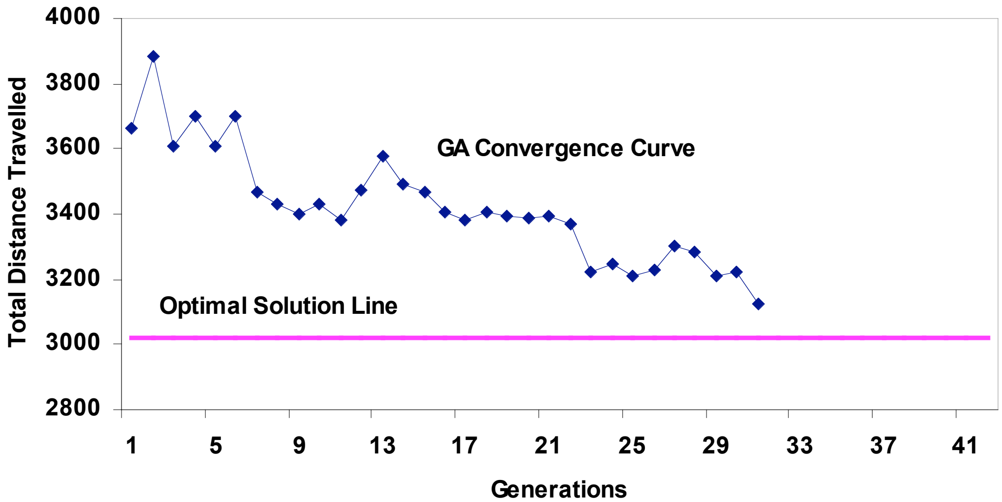


figure 8:GA convergence for period 4

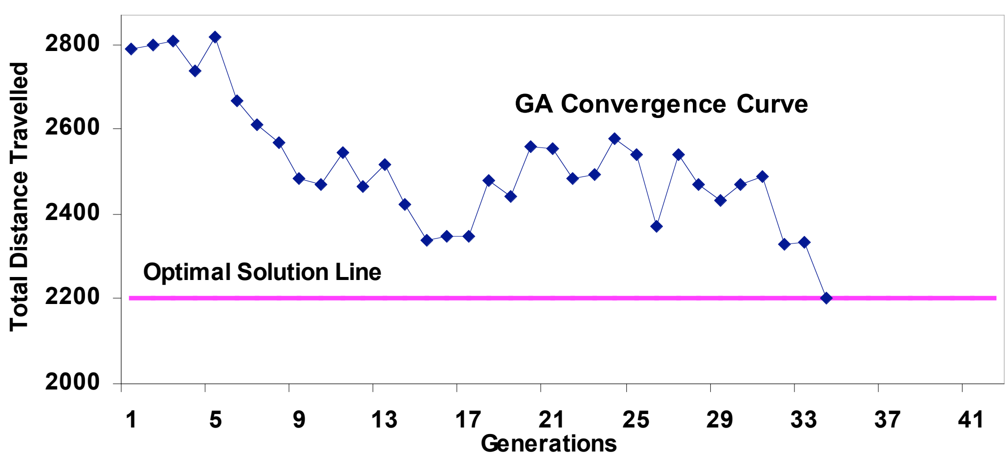


figure 9:GA convergence for period 5

6. Conclusion

In this paper, a model is developed for machine assignment problem in cellular manufacturing layout with the objective of minimizing the total traveling distance. Genetic Algorithm is used to solve the problem. GA works well for this machine assignment problem. The computational results show that the GA is capable of obtaining near optimal (99%) solution of the test problem and out performs other techniques. GA is simple and easily modifies the parameters according to the changes and also it has relatively low computational time. In this paper, all the machine locations are considered to be of equal size. This work may be extended by considering the machine locations of unequal size.

References

Bazargan-Lari.M.(1999) Layout designs in cellular manufacturing. *European Journal of Operational Research; 112*: 258-272

Bazargan-lari.M., Kaebernick.H. and Harraf.A.(2000) Cell formation and layout designs in a cellular manufacturing environment- a case study. *International Journal of Production Research; 38(7)*: 1689-1709.

Castillo.I. and Peters.BA.(2003) An extended distance-based facility layout problem. . *International Journal of Production Research; 41(11):* 2451-2479.

Chan.W.M., Chan.C.Y. and Ip.W.H.(2002) A heuristic algorithm for machine assignment in cellular layout. . *Computers & industrial Engineering; 44*: 49-73.

Chiang.W.C. and Chiang.C.(1998) Intelligent local search strategies for solving facility layout problems with the quadratic assignment problem formulation. *European Journal of Operational Research; 106:* 457-488

Chiang.W.C. and Kouvelis.P.(1996) An improved tabu search heuristic for solving facility layout problems. *International Journal of Production Research; 34(9)*: 2565-2585

Dweiri.F. and Meier.F.A.(1996) Application of fuzzy decision-making in facilities layout planning. . *International Journal of Production Research; 34(11):* 3207-3225.

Hassan.M.M.D,(1994) Machine layout problem in modern manufacturing facilities. . *International Journal of Production Research; 32(11):* 2559-2584.

Hassan.M.M.D.(1995) Layout design in group technology manufacturing. *International Journal of production Economics; 38*: 173- 188

Heragu.SS. and Kusiak.A,(1990)Machine layout: an optimization and knowledge-based approach. . *International Journal of Production Research; 28(4):* 615-635

Jajodia.S., minis.I., Harhalakis.G. and Proth.JM., (1992)CLASS: Computerized Layout Solutions using Simulated annealing. *International Journal of Production Research; 30(1*): 95-108.

Kalyanmoy Deb. (2002) *Optimization for engineering design: Algorithms and examples*. Prentice hall of India, New Delhi.

Meller RD, Gau KY,(1996) The Facility Layout Problem: Recent and Emerging Trends and Perspectives. *Journal of Manufacturing Systems; 15(5)*: 351-366

Salum.L.(2000) the cellular manufacturing layout problem. *International Journal of Production Research; 38(5)*: 1053-1069

Souiah.A,(1995) Simulated Annealing for manufacturing systems layout design. *European Journal of Operational Research*; 82: 592-614

Tam.K.Y. and Li.S.G.(1991) A hierarchical approach to the facility layout problem. *International Journal of Production Research; 29:* 165-184.

Tanchoco.J.M.A. and Lee.A.C (1999) Cellular machine layout based on the Segmented Flow Topology. *International Journal of Production Research; 37(5):* 1041-1062.

Tang.C. and Abdel-Malek.L.L.(1996) A frame work for hierarchical interactive generation of cellular layout. *International Journal of Production Research; 34(8)*: 2133-2162.

Urban.TL., chiang.WC. and Russell.RA.(2000) The integrated machine allocation and layout problem. *International Journal of Production Research. ; 38(13*): 2911-2930

Wang.T.Y. , Lin.H.C. and Wu.K.B. (1998) An improved simulated annealing for facility layout problems in cellular manufacturing systems. *Computers ind. Engng; 34(2)*: 309-319

Welgama.P.S. and Gibson.P.R.(1996) An integrated methodology for automating the determination of layout and material handling system. *International Journal of Production Research; 34(8*): 2247-226

Yaman.R, Gethin.D.T. and Clarke.M.J. (1993) An effective sorting method for facility layout constructions. *International Journal of Production Research; 31(2)*: 413-427.

A PIECEWISE LINEAR APPROXIMATION PROCEDURE FOR L_p NORM CURVE FITTING

Michael G. Sklar,Rutgers University, USA
Ronald D. Armstrong,Rutgers University, USA

ABSTRACT

There are numerous applications of fitting a linear function to a set of observed points using an objective of minimizing the norm of the error vector in some sense. This paper is concerned with the L_p norm criterion, which for p greater than 1 is equivalent to an unconstrained nonlinear convex minimization problem. Many standard *nonlinear* algorithms exist to solve this nonlinear problem, but we propose an alternative approach; a piecewise *linear* approximation of a linearly constrained equivalent problem. For computational efficiency the algorithm makes use of an extension of the "multi-vertex" pivot rules which have been effectively employed in L_1 norm (Least Absolute Value) regression. Computational results of the algorithm testing are presented, together with a simple comparison to a published algorithm for solving unconstrained, nonlinear minimization problems.

Keywords

L_p Norm, Linear Approximation, Separable Programming, Multi-Vertex Pivots, Convex Minimization, Regression

PURPOSE

The use of a normed estimation criterion when estimating the parameters in linear regression is widespread, simple, and popular. Least squares estimation is the most common of these methods - it may be called the L_2 norm estimate, where the "2" indicates it is the sum of the *squared* errors which is being minimized: minimize $\Sigma(\text{errors})^2$. Similarly, the frequently used estimation procedure Least Absolute Values (LAV) may be called the L_1 norm, where the sum of the *absolute* errors is being minimized. Both of these methods represent linear functions to find optimal parameter estimates, but other, nonlinear, methods may be useful in estimation. That is, minimizing the sum of errors raised to some power besides 1 or 2 can be useful, such as the $L_{1.5}$ norm, where we minimize $\Sigma(\text{errors})^{1.5}$. In this paper we propose an algorithm to approximate the nonlinear function by making adjacent pieces of the function linear, and using a modified Linear Programming algorithm to solve the resulting piecewise approximation. In other words, for normed estimation other than least squares and least absolute values, the errors or residuals are nonlinear, and our algorithm linearizes the residuals, and then uses linear programming to minimize these residual pieces.

There is often an advantage to using some norm ($L_{1.5}$ for instance) in place of least squares (L_2 norm) or LAV (L_1 norm). Least squares is analogous to the mean in the scalar case, and LAV is analogous to the median. We propose a more-general and flexible approach of minimizing the sum of errors raised to any power ≥ 1. This compromise between the mean and the median may be useful for exploratory data analysis and for estimation.

INTRODUCTION

The L_p norm curve-fitting criterion, for $p > 1$, has been effectively modified for certain cases. Specializations of the simplex algorithm, when $p = 1$ or for $p \to \infty$, have been shown to provide significant savings in computer time and storage space, compared to the standard simplex algorithm. The piecewise linear approximation algorithm described here utilizes an extension of the "multi-vertex" pivot rules used by efficient L_1 norm algorithms. A multi-vertex pivot may be described as passing through several vertices (extreme point solutions) in a single pivot, where the basis inverse is not updated until the pivot is complete.

The curve fitting model of interest here is

$$\underline{y} = X\underline{\beta} + \underline{r}, \quad (1)$$

where X is the n by m matrix of n observations of the m factors affecting the dependent variable Y, $\underline{\beta}$ is the vector of the m parameters to estimate, $\underline{y}$ gives the vector of observed values of Y and $\underline{r}$ is an n-vector of random errors. The goal is to fit the hyperplane $X\underline{\beta}$ to minimize the error vector $\underline{r}$ in some sense. Writing the pth norm as

$$L_p = \| \underline{r} \|_p = \left[\sum_{i=1}^{n} | r_i |^p \right]^{1/p}$$

we wish to minimize the pth norm for $p \geq 1$. The norm is a strictly convex function for $p > 1$, so that a nonlinear algorithm is typically used to solve the minimization. When p is 1, 2, or for $p \to \infty$, an exact linear programming equivalent can be written. Kennedy and Gentle (1977) in their book discuss L_∞ norm estimation, where they provide an extensive bibliography.

The procedure proposed here is based on the simplex method, and it solves the convex problem by linearizing pieces of the convex graph and operating on these linear pieces to obtain the solution. This algorithm, then, provides a piecewise, linear approximation of the convex function.

Minimizing the norm of the residual vector $\underline{r}$ for some value of p besides p = 1 (least absolute values) and p = 2 (least squares) may offer a useful compromise for curve fitting. The robust properties of least absolute value (LAV or L_1 norm) estimation are well known: LAV provides protection against outlying values, in the same sense that the median does in the scalar case. This outlier protection is at the cost of increased estimator variance compared to the variance from using least squares (L_2 norm) when the errors are, in fact, normal. As a compromise, then, $L_{1.5}$ might be appropriate, so that some of the outlier protection of the L_1 norm is obtained, but with smaller estimator variance.

Two modifications are described which significantly improve the efficiency of our algorithm over the standard linear programming approach: employing a reduced basis (with significantly fewer representation updates per pivot), and using a multi-pivot approach to skip pivots.

THE FORMULATION

The formulation of the piecewise linear approximation will be described below. Conceptually, each residual r_i may be viewed in terms of a convex graph (see Figure 1) with possible values of r_i plotted along the abscissa and the corresponding absolute residual raised to the pth power, $|r_i|^p$, along the ordinate. The resulting convex curve for r_i is then segmented into a number of linear pieces, of width Δ (which is chosen by the user of the algorithm). Under this scheme, values to the right of the ordinate pertain to a positive residual. The segment width, Δ, (see Figure 1) determines how many segments or pieces are needed to represent the residual r_i - either positive or negative residual. From the simplex algorithm viewpoint, each *segment* making up the residual r_i will be represented by a decision variable, d_{ij}^*: d_{ij}^+ if $r_i > 0$, and d_{ij}^- if $r_i <$ 0, so that

$$|r_i| = \sum_{i=1}^{k(i)} d_{ij}^*$$, where k(i) is defined below.

(Throughout the paper, we use "*" to denote that d_{ij}^* pertains to either d_{ij}^+ or d_{ij}^- .)

Several terms are defined below to assist in the description of the algorithm.

Joint Point: If a residual is at a *joint point*, each d_{ij}^* piece making up the residual is at its bound.

In Figure 1, the indicated residual is represented by 4 variables: the first three are at their bound of Δ (thus each is at a joint point), and the last one, d_{i4}^+ , is less than its bound. Thus the indicated residual is not at a joint point, since not all of its variables are at their lower or upper bound (0 or Δ in this case). If the residual r_i *were* at a joint point, all of its variables would be at their bound, and there would be no need for a basic variable $d_{i4}^+ < \Delta$.

f(μ), $\mu = 1,...,m$ is an index set which indicates which rows of X make up the basis matrix **B,** so $f(\mu) = s$ indicates that the μ-th row of **B** is the original row s;

k(i), $i = 1,...,n$ is an index set which indicates which segment the ith residual is at, so if $k(i) = t$, then the ith residual is at segment t, and $|r_i| = \sum_{j=1}^{t} d_{ij}^{*}$. If the residual r_i is **at** a joint point, then $k(i) = t$ implies that all the d_{ij}^{*} for $j = 1,...,t$ are at the upper bound of Δ, and none of these variables would be basic in the full structure. If r_i is **between** joint points, then $k(i) = t$ means that all the d_{ij}^{*} for those j *before* t are at their upper bound, and the final piece d_{it}^{*} is between its bounds and would be basic in the full structure.

K(i) is the number of pieces used by the linear programming model to make up the residual r_i.

The graph of Figure 1 indicates that this particular residual r_i takes four variables, k(i)=4, to describe it:

$$r_i = (d_{i1}^{+}=\Delta) + (d_{i2}^{+}=\Delta) + (d_{i3}^{+}=\Delta) + (0 < d_{i4}^{+} < \Delta).$$

Since the last piece $d_{i4}^{+} > 0$ but not at its bound, r_i is not at a joint point. Using the simplex method for bounded variables, r_i would be described by three variables which are non-basic at their upper bound Δ, and one basic variable less than Δ. If $d_{i,k(i)}^{*}$ is the basic variable, all variables before it ($d_{ij}^{*} : j < k(i)$) are at their upper bound of Δ, and all variables which come after $d_{i,k(i)}^{*}$ ($d_{ij}^{*} : j > k(i)$) are at their lower bound, which is zero.

pth norm: $L_p = \| \underline{r} \|_p = \left[\sum_{i=1}^{n} | r_i |^p \right]^{1/p}$

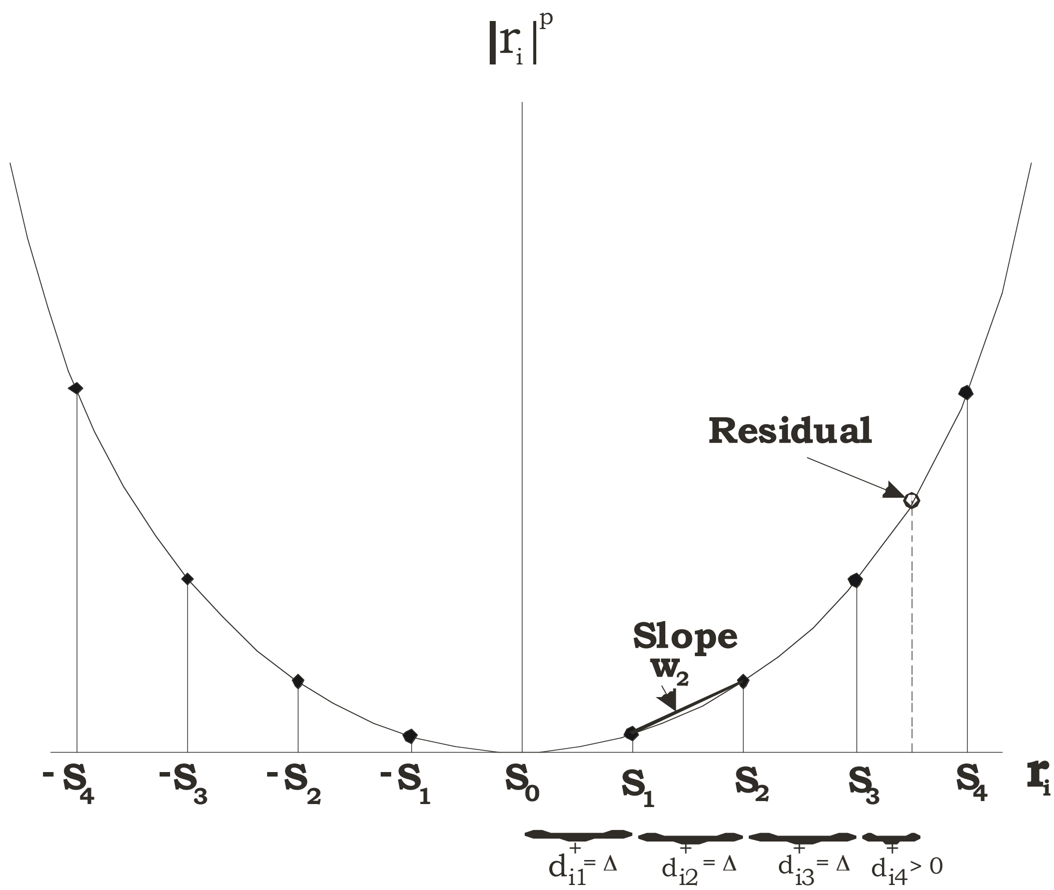


Figure 1. The graph of residual values versus pth power residual values, where each solid vertical line represents a joint point for this residual (where the associated variable is at its upper bound). For this example, four variables are required to represent this specific residual r_i: three variables at their upper bound of Δ, and a fourth non-zero variable which is less than its upper bound. The dotted line indicates the position of the residual.

Since a portion of the convex graph is linearized (as shown in Figure 1), the slope may be determined. We define w_k as the slope for the kth linearized segment, where

$$w_k = \frac{(S_k)^p - (S_{k-1})^p}{(S_k - S_{k-1})} = \frac{(S_k)^p - (S_{k-1})^p}{\Delta}.$$

The linear programming formulation then becomes

$$\text{Min} \sum_{i=1}^{n} \sum_{k=1}^{K(i)} (w_k\, d_{ik}^{+} + w_k\, d_{ik}^{-}) \tag{2}$$

$$\text{Subject to} \sum_{j=1}^{m} X_{ij} b_j + \sum_{k=1}^{K(i)} d_{ik}^{+} - \sum_{k=1}^{K(i)} d_{ik}^{-} = y_i, \quad i = 1,2,...,n.$$

$$d_{ij}^{-} \geq 0,\ d_{ij}^{+} \geq 0\ ,$$

$$b_j \text{ unrestricted.}$$

In theory, there is no limit on the value of K(i), which equals the number of pieces used to represent a residual. The number of variables needed to describe r_i is a function of the size of r_i and of Δ (the bound, or interval width, or the maximum value of each variable d_{ij}^{*} making up the residual), but we will assume K(i) is finite. (The computer implementation of the algorithm does not limit the value of K(i) and segments are not defined explicitly, but are generated as needed.)

In matrix terms,

$$\text{Min} \quad \underline{w}\,\underline{d}^{+} + \underline{w}\,\underline{d}^{-} \tag{3}$$

$$\text{s.t.} \quad X\underline{b} + H^{+}\underline{d}^{+} + H^{-}\underline{d}^{-} = \underline{y}$$

$$\underline{0} \leq \underline{d}^{+} \leq \underline{1}\,\Delta$$

$$\underline{0} \leq \underline{d}^{-} \leq \underline{1}\,\Delta$$

$$\underline{b} \text{ unrestricted.}$$

Note that the bounding restrictions on $\underline{d}^{+}$ and $\underline{d}^{-}$ will not be treated as explicit constraints in the operation of the simplex algorithm; the simplex rules will be altered to take these bounds into account. Let H_i^* be the column for d_{ij}^{*}, where we drop a subscript and use H_i^* instead of H_{ij}^*, since the column is the same for every j. (That is, H_i^+ is the column for d_{ij}^+ for every j and similarly for H_i^-.) Then, if d_{ij}^* is basic, the column $H_i^* = u_i^*$ is in the basis **F** (described below), where u_{i^*} is used to denote an n-vector with zeros in each position except for a +1 in the ith position if * = "+", or a -1 in the ith position if * = "−" . After reordering, these ±1 for basic d_{ij}^* make up the diagonal matrix J. It follows from the theory of separable convex programming that only those nonbasic d_{ij}^* with their ±1 in the first m rows (after reordering) will be potential candidates to enter the basis, and the column associated with each ±1 is of the form $H_i^* = u_i^* = \begin{bmatrix} e_i^* \\ \underline{0} \end{bmatrix}$, where e_i^* is an m-vector defined in the same way as the n-vector u_i^* ; e_i^* is an m-vector with a ±1 in the ith position.

When X is partitioned as $X = \begin{bmatrix} B & \text{m by m} \\ R & \text{n-m by m} \end{bmatrix}$, where **B** has full row rank, problem (1) may be written as

$$\begin{array}{lllllll} \text{Min} & & \underline{w}\, d_B^+ & + \underline{w}\, d_R^+ & + \underline{w}\, d_B^- & + \underline{w}\, d_R^- & \\ \text{s.t.} & B\underline{b} & + H_B^+ d_B^+ & & + H_B^- \underline{d}_B^- & & = Y_B \\ & R\underline{b} & & + H_R^+ \underline{d}_R^+ & & + H_R^- \underline{d}_R^- & = Y_R \end{array} \qquad (4)$$

$$0 \le \underline{d}^+ \le \underline{1}\,\Delta$$

$$0 \le \underline{d}^- \le \underline{1}\,\Delta$$

$\underline{b}$ unrestricted.

A basis **F** for this problem may be partitioned as follows:

$$\mathbf{F}\ (\text{n by n}) = \begin{bmatrix} B & (\text{m by m}) & : & 0 & (\text{m by n-m}) \\ R & (\text{n-m by m}) & : & J & (\text{n-m by n-m}) \end{bmatrix},$$

Matrix X : $\underline{d}^-$ *and* $\underline{d}^+$

Without loss of generality, assume the columns of **F** have been arranged so that the first m columns of **F** (which is the matrix X) are associated with the $\underline{b}$ variables. (The $\underline{b}$ variables, being unrestricted, will be placed in the initial basis **F** and remain in the basis throughout the algorithm.) The remaining n-m columns of **F** pertain to $\underline{d}^+$ and $\underline{d}^-$, and the rows have been interchanged so that **J** is a diagonal matrix. It is the m by m matrix **B** which is of special interest, since it will be used as a working or reduced basis instead of the n by n basis **F**. We will distinguish between two structures: The "full structure," which uses the n by n basis **F**, and which has n rows (the upper bound restrictions on the d^*_{ij} are handled implicitly, so they don't increase the basis size); and the "reduced structure," which utilizes the m by m basis **B**, resulting in significantly fewer representation updates.

The distinction is important because the viewpoint using the basis **B** is quite different from the viewpoint using **F**. In the full structure (using **F**), a pivot causes the column for a nonbasic d^*_{ij} to enter **F**, which results in a new column of **F** with the ±1 replacing an entry in the $\underline{\mathbf{0}}$ matrix. Further, the column which leaves **F** had a ±1 in the matrix **J**, which is replaced by a zero. To maintain the partitioning of **F**, the row with the new ±1 in the $\underline{\mathbf{0}}$ matrix is interchanged with the row which has the new "0" in the last n-m rows (in **J**). Therefore, a pivot from the viewpoint of the full structure using **F** is tantamount to replacing a column of **F** and then interchanging two rows of the newly formed full basis to preserve the specified structure. Said another way, since the parameter estimates $\underline{b}$ will never leave the basis, an incoming column will replace an entry in the ($\underline{\mathbf{0}}$,**J**) part of the basis **F**, and a row from **B** and a row from **R** will be exchanged.

The column entering the full basis **F** places a ±1 into the $\underline{\mathbf{0}}$ matrix, but to maintain the partitioning of **F**, that row in $\underline{\mathbf{0}}$ with the new ±1 is moved down so the ±1 is now in **J** . This movement from the reduced basis perspective causes a row to leave **B** and move to **R**. Thus a ***column entering*** the full basis **F** results in a ***row leaving*** the reduced basis **B**. Because for efficiency we focus on the reduced basis **B**, the pivot computations are in terms of **B**. But the reduced cost values computed to determine the ***leaving row*** of **B** would be the exact same values we would compute to determine an ***entering column*** of **F**, and the minimum ratio values computed to determine the ***entering row*** of **B** would be the exact same values we would compute to determine a ***leaving column*** of **F**. Thus the algorithm can be viewed from either the full or reduced structure, although the implementation (as well as the computation) is simplified with the reduced structure.

The statement that a residual r_i is at a "**joint point**" means that *all* the variables which make up an r_i^* are at their bound, so that (in the absence of degeneracy) none of the d_{ij}^* which make up an r_i at a joint point is basic in the full structure; each is nonbasic at a bound. That is, the residual is exactly some multiple of Δ so all of its variables (with value Δ) will be non-basic - then that r_i is associated with the rows of (**B**,**0̲**). Conversely, r_i is **not** at a joint point if the last of its variables is < Δ so it would have to be basic in **F** - then that r_i is associated with the rows of (**R**,**J**) .

The general procedure is to measure the effect (using the reduced costs) of increasing or decreasing a d_{ij}^* associated with a row of (**B**,**0̲**). If the effect of changing a d_{ij}^* is to reduce the objective value, then this d_{ij}^* changes in value until one residual r_i is changed sufficiently that a basic d_{ij}^* associated with a row of (**R**,**J**) leaves the basis **F,** or d_{ij}^* itself moves to a bound. Thus, the residual associated with the d_{ij}^* leaving the basis moves to a joint point (so that all of the d_{ij}^* which make up the residual are at a bound).

THE REDUCED BASIS B

We propose to use the reduced basis **B** rather than the full basis **F** for the simplex computations. Note that while the number of nonbasic d_{ij}^* may be quite large, depending on the size of Δ, we need determine reduced costs for only *2m* of these: For each of the *m* residuals at a joint point, one variable may increase or an adjacent variable may decrease and those are the only ones we need to consider. The coefficient matrix for the original problem described in (3) may be denoted as $\left(\begin{bmatrix} B \\ R \end{bmatrix} H^+, H^- \right)$, for which a basis **F** has been specified earlier: $\mathbf{F} = \begin{bmatrix} B & \underline{0} \\ R & J \end{bmatrix}$. The inverse is $\mathbf{F}^{-1} = \begin{bmatrix} B^{-1} & \underline{0} \\ -JRB^{-1} & J \end{bmatrix}$.

The slope of the linearized portion is denoted w_{ij}, which pertains to the jth linearized segment for the ith residual. However, the slope depends only on the segment (j) of the convex curve, and not on the residual (i). For simplification, then, the slope will be denoted w_j, a scalar. Collecting these slopes, $\underline{w}_R$ is the vector of slopes associated with those d_{ij}^* which are basic in the full structure and $\underline{w}_B = \underline{0}$ is the vector of objective function coefficients associated with the parameter estimates $\underline{b}$.

The reduced cost for the residual r_i is computed using the full basis **F** by $\bar{c}_{ij} = w_j - C_F F^{-1} H_i^*$. In terms of the reduced basis **B,** the identical reduced costs are

$$\bar{c}_{ij} = w_j - C_F F^{-1} H_i^* = w_j - (\underline{w}_B, \underline{w}_R) \begin{bmatrix} B^{-1} & \underline{0} \\ -JRB^{-1} & J \end{bmatrix} \begin{bmatrix} e_i^* \\ \underline{0} \end{bmatrix} \Rightarrow \bar{c}_{ij} = w_j - (-w_R JRB^{-1})_i.$$

In selecting the row to leave **B,** we consider changing the value of those residuals at a joint point: $d^*_{f(i),k(i)}$. Suppose r_i is at the t-th joint point, so t = k(i). Then either of two adjacent variables may be changed to modify r_i: d^*_{it} may decrease from its upper bound of Δ or $d^*_{i,t+1}$ may increase. (A special case occurs when r_i is at the ***zero*** joint point; t = 0. Here, either of the adjacent variables d^+_{i1} or d^-_{i1} may increase.) Whether r_i is at a zero joint point or not, which of the two adjacent variables to change will be determined by reduced costs - the greatest marginal improvement in the objective function.

We define a variable

$$\text{Idelta}(d^*_{i,k(i)}) = \begin{cases} 1 \text{ if } r_i \text{ is increasing (so } d^*_{i,k(i)+1} \text{ is increasing from its lower bound)} \\ -1 \text{ if } r_i \text{ is decreasing (so } d^*_{i,k(i)} \text{ is decreasing from its upper bound of } \Delta) \end{cases}$$

Using Idelta enables the same test to be used for those variables increasing and for those decreasing. We select row s to leave **B,** where row s corresponds to the $d^*_{s,k(i)}$ or $d^*_{s,k(i)+1}$ satisfying the reduced cost criterion

$$\max \left\{ \bar{c}_{f(i),k(i)} * [\text{Idelta}(d^*_{f(i),k(i)})] < 0 \right\},$$

$$\text{or} \left\{ \bar{c}_{f(i),k(i)+1} * [\text{Idelta}(d^*_{f(i),k(i)+1})] < 0 \right\},$$

so that candidates for the leaving row are always those with negative reduced costs. From the perspective

of the full structure, the selected variable, $d^*_{s,k(i)}$ or $d^*_{s,k(i)+1}$, will enter the basis **F.**

SELECTING THE ROW TO MOVE FROM R TO B, WITHOUT MULTIPLE-VERTEX PIVOTS

A minimum ratio rule is used to select the column to leave **F,** which in effect moves a row of **R** into **B.** In the *full structure* it is necessary to compute the representations of the right hand side vector and for the incoming column vector for the variable d^*_{st}, in terms of the basis **F**. Let e^*_s represent an m-dimensional vector with zeros for every entry except a ±1 in the s-th position.

$H^*_s = \begin{bmatrix} e^*_s \\ \underline{0} \end{bmatrix}$, so $\bar{a}_s = F^{-1}H^*_s$; and the new representation of the right hand side vector is $y = F^{-1}\hat{y}$, where $\hat{y}$ is the original right hand side adjusted by the nonbasic variables at their upper bounds. The representations for the reduced problem may be similarly computed in terms of the reduced basis **B.** Here is the representation for the nonbasic columns (excluding the deviation variables) where the full basis **F** is written in terms of the reduced basis **B**:

$$\bar{y} = F^{-1}\hat{y} = \begin{bmatrix} B^{-1} & \underline{0} \\ -JRB^{-1} & J \end{bmatrix} \begin{bmatrix} y_B \\ y_R \end{bmatrix} = \begin{bmatrix} B^{-1}y_B & \text{for the rows of } \mathbf{B} \\ -JRB^{-1}y_B + Jy_R & \text{for the rows of } \mathbf{R} \end{bmatrix} .$$

Similarly, here is the representation for the nonbasic deviational variable columns where the full basis **F** is written in terms of the reduced basis **B**:

$$\bar{a}_S = F^{-1} H^*_S = \begin{bmatrix} B^{-1} & \underline{0} \\ -JRB^{-1} & J \end{bmatrix} \begin{bmatrix} e^*_s \\ \underline{O} \end{bmatrix} = \begin{bmatrix} B^{-1}e^*_s & \text{for the rows of } \mathbf{B} \\ -JRB^{-1}e^*_s & \text{for the rows of } \mathbf{R} \end{bmatrix} .$$

In the reduced problem, it is a row of **R** which is selected to enter (or move to) **B,** so the row chosen to leave **R** and enter **B** corresponds to the minimum ratio

$$\min \left\{ \frac{-JRB^{-1}y_B + Jy_R}{-JRB^{-1}e^*_i} \right\}$$ for $1 \le i \le m$. Because of the upper bound restrictions, there are the usual cases to consider in selecting the minimum, which will be described both from the full and reduced basis viewpoints: the minimum occurs for one of the following cases:

(1) a d^*_{ij} leaving **F** decreases to its lower bound;

(2) a d^*_{ij} leaving **F** increases to its upper bound Δ;

(3) a d^*_{ij} entering **F** decreases to its lower bound (so a pivot does not occur); and

(4) a d^*_{ij} entering **F** increases to its upper bound of Δ (no pivot occurs).

To summarize the relationship between the full structure and the reduced structure: The reduced cost $\bar{c}_{ij}$ which selects the **column to enter F** in the *full structure* is precisely the reduced cost which identifies the **row to leave B** in the *reduced structure* sense; the minimum ratio associated with the column for d^*_{ij} which is selected to leave **F** is exactly the ratio which identifies the row to enter **B.** Therefore, in the algorithm (and in the accompanying computer code), the computations are carried out in terms of the reduced basis **B,** where a row is selected to leave **B** (on the basis of reduced costs) and a row is selected to enter **B** (using the minimum ratio).

THE MULTIPLE-VERTEX (M-V) PIVOT CONCEPT

Being able to essentially skip pivots (basis updates) can result in a more efficient algorithm. The strategy of a multiple-vertex pivot (in effect accomplishing several pivots at one time with very little additional computational effort) will be described from both the full and reduced structure viewpoints. The key observation is that, considering the residual $r_i = \sum_{i=1}^{k(i)} d^*_{ij}$, the column for d^*_{ij} is the same for every j. Therefore, from the full structure perspective, if a current basic variable d^*_{ij} is replaced by an adjacent variable, $d^*_{i,j+1}$, the basis **F** will not change, since both variables have the same column. (The columns could differ by a sign, of course, if $d^+_{i,1}$ is replaced by $d^-_{i,1}$ or vice versa, in which instance the residual r_i would have changed sign.)

The method for M-V pivots will be as follows (from the viewpoint of the full structure): Suppose $r_i = \sum_{j=1}^{t} d^+_{ij}$ is at a *joint point*; no d^+_{ij} is basic, so that $d^+_{ij} = \Delta$, j = 1,2,...,t, and $d^+_{ij} = 0$, j > t. Suppose the gradient vector of reduced costs indicates that decreasing r_i will improve the objective function. Then r_i's last variable d^+_{it} is selected to enter the basis, decreasing from its upper bound of Δ. Using the minimum ratio computations, the qth basic variable, $d^*_{f(q),k(d(q))}$, is selected to leave. More specifically, suppose that the qth basic variable is d^+_{uv} so f(q) = u (the u-th row of **B** is the original row q), and k(u) = v (the u-th residual is at segment v).

Since r_u is at a joint point, $r_u = \sum_{j=1}^{v} d^+_{uj}$, where $d^+_{uj} = \Delta$, for j = 1,...,v, and $d^+_{uj} = 0$, for j > v. (That is, each d^+_{uj} making up the residual r_u is at its upper bound Δ.)

To summarize, let d^+_{it} be selected to enter the basis, let d^+_{uv} be selected to leave the basis, and both r_i and r_u are decreasing. Without a M-V pivot, the column for d^+_{it} would replace the column for d^+_{uv} in **F,** and an *explicit pivot would take place*. In other terms, the residual r_i would decrease from a joint point to some interval position (d^+_{it} would become basic) while r_u would decrease to a joint point (so that each piece making up r_u, d^+_{uj}, is nonbasic for every j). In order to utilize the M-V pivot strategy, we follow the logic described below. Suppose the movement is even more than "allowed" by the standard ratio test, so that r_u decreases not just to the next joint point (no basic variables), but beyond that. Then the adjacent variable, $d^+_{u,v-1}$, becomes basic in place of d^+_{uv}. If the entering d^+_{it} would still be a candidate to enter the basis (determined by looking at the new gradient if this extra movement is allowed) then the "M-V pivot" is allowed; that is, d^+_{uv} decreases to its lower bound, but r_u decreases even more because the adjacent variable $d^+_{u,v-1}$ decreases from its upper bound of Δ, so $d^+_{u,v-1}$ becomes the qth basic variable, and the original entering variable d^+_{it} is still a candidate to enter in place of another variable.

In this situation, the basis **F** doesn't change, but the basic variable values and the reduced costs change. New ratios are computed to select the leaving variable, and the pivoting process continues. A similar approach is used when the minimum ratio occurs for the entering variable d^+_{it} decreasing to its bound. If d^+_{it} decreases to its bound, would the adjacent variable, $d^+_{i,t-1}$, be a candidate to enter the basis? If so, the M-V pivot is allowed, where d^+_{it} goes to zero, $d^+_{i,t-1}$ is a candidate to decrease from its upper bound of Δ by entering the basis, and the required updates are made. The M-V pivot approach from the viewpoint of the reduced basis **B** follows the same logic.

COMPUTATIONAL RESULTS AND CONCLUSIONS

Computational results are provided using this special-purpose algorithm for solving a piecewise linear approximation of an unconstrained, nonlinear convex minimization problem.

The results are compared to the solution times obtained from a published nonlinear algorithm by Shanno and Phua (1980).

A variety of problems were considered, although only a small subset are reported here. A number of combinations of m (the number of parameters (β) to estimate), n (the number of observations), p (p is the power pertaining the pth norm), and the three distributions were tested. The tested values of m were {2, 3, 6}, the choices for n were {50, 100, 200, 500}, and the distributions used were Uniform($-\sqrt{3},\sqrt{3}$), Normal(0,1), and Double Exponential with zero mean. A variety of choices of p were tested, where the pth norm may be characterized by $\|\underline{y} - X\underline{b}\|_p = \left[\sum_{i \in I} |(y - X\underline{b}_i|^p)\right]^{1/p}$, $i \in I = \{1,2,\ldots,n\}$,

and p≥1); the IMSL Library was used to generate the X and y data. For testing we utilized numerous choices of the segment width. The tradeoff is that with a smaller width we have more accuracy but many more segments (variables) to process. We utilized widths of 0.05, 0.01, 0.1, 0.2, and so on. We used an advanced start with a first pass having width 1.0, then switched to a smaller width such as 0.1 for the fine tuning. We also set up an automatic procedure which began with length equal to some small value, then successively changed the length by computing "new length" = length/10 until the difference in results was less than some small tolerance. We leave for future work the determination of the best choice of width.

Our piecewise approximation LSUBP performed better than the NONLINEAR algorithm as the values of p approached one. The nonlinear algorithm, since it uses derivatives to find a direction, has more trouble finding the derivatives as the curve becomes more nearly linear. As the value of p increased, NONLINEAR performed better. Also, LSUBP did better when the distribution of the errors was exponential, and this is the error distribution where it would be particularly advisable to use a norm such as p=1.1 . Because the multi-vertex concept has worked very well for simplex-based L_1 norm algorithms (Armstrong, Frome, and Kung 1978) and L_∞ norm algorithms (Armstrong and Kung 1979), we felt the simplex-based piecewise linear approximation method for L_p norm estimation would also prove efficient. We believe that the presentation of this approach serves a useful purpose for several reasons: it answers a legitimate research question regarding the efficacy of using a piecewise linear approach; it may lead to additional exploration of this approach, and it may provide a useful teaching foundation.

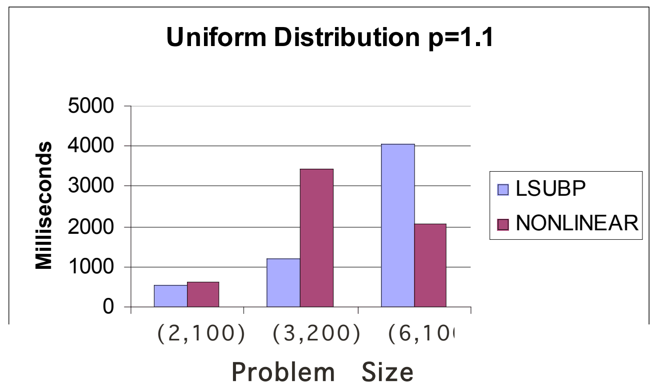
Figures: A subset (for p=1.1) of computational results for our special-purpose, simplex-based algorithm (LSUBP) and a nonlinear algorithm (NONLINEAR), for three underlying distributions. The reported times are in milliseconds using a workstation.
The 3 problem sizes considered were

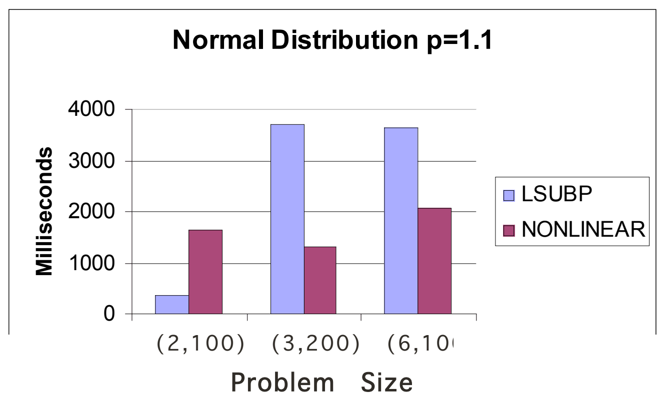
number of parameters	**number of observations**
2	100
3	200
6	100

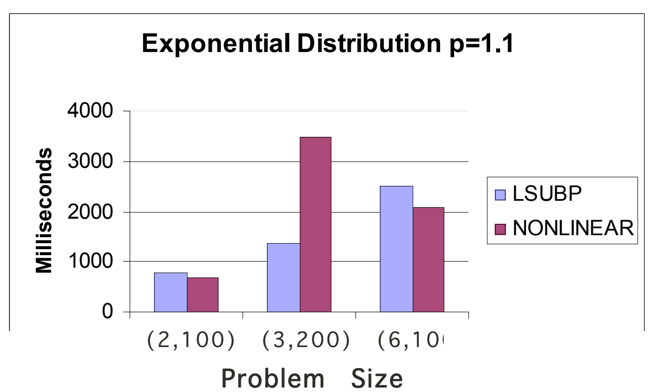
p is the norm index:

$$\| r_i \|_p = \left[\sum_{i \varepsilon I} | (\underline{y} - X\underline{b}) |^p \right]^{1/p},$$

$i \varepsilon I = \{1,2,...,n\}$, and $p \geq 1$.







REFERENCES

Armstrong, R. D., Frome, E. L., and Kung, D. S. (1978). A revised simplex algorithm for the absolute deviation curve-fitting problem. *Communications in Statistics B(8)*,175-190.

Armstrong, R. D. and Kung, D. S. (1979). Minimax estimates for a linear multiple regression problem. *Applied Statistics 28*, 93-100.

Ekblom, H., and Henriksson, S. (1969). L_p-Criteria for the Estimation of Location Parameters. *Siam J. Appl. Math 17(6)*, 1130-1141.

Forsythe, A. B. (1972). Robust Estimation of Straight Line Regression Coefficients by Minimizing pth Power Deviations. *Techonometrics* 14, 159-166.

Gentle, J. E. (1977). Least Absolute Values Estimation: An Introduction. *Commun. Statist. –Simul. Computa. B6(4)*, 313-328.

Gentle, J. E., Kennedy, W. J. and Sposito, V. A. (1977). On Least Absolute Deviations Estimation. *Comm. Stat. A(6)*, 839-845 .

Gentle, J. E., Kennedy, W. J., and Sposito, V. A. (1977). Properties of L_1 Estimators. *Math. Prog. 12*, 138-140.

IMSL Statistical Library, 14141 Southwest Freeway, Suite 3000, Sugar Land, Texas, 77478-3498

Kennedy, W. J. Jr. and Gentle, J. E. (1980). Statistical Computing, Marcel Dekker, Inc., New York.

Money, A. H., Affleck-Graves, J. F., Hart, M. L., and Barr, G. D. I. (1982), The Linear Regression Model: L_p Norm Estimation and the choice of p. *Comm. Statist. -Simula. Computa. 11(1)*, 89-109.

Rice, J. R., and White, J. S. (1964). Norms for Smoothing and Estimation. *Siam Review 6(3)*, 243-256.

Shanno, D. F. and Phua, K. H. (1980). Remark on algorithm 500: minimization of unconstrained multivariate functions. *Transactions on Mathematical Software* 6, 618-622.

Sklar, M. G. and Armstrong, R. D. (1988) Extensions to a Best Subset Algorithm for Least Absolute Value Estimation, *American Journal of Mathematical and Management Sciences, 8(1&2),* 1-58 .

An Integrated Information System for Customer Requirement Analysis

Ying-Shen Juang, Chung-Hua University, Taiwan, R.O.C.

Shui-Shun Lin, National Chinyi Institute of Technology, Taiwan, R.O.C.

Hsing-Pei Kao, National Central University, Taiwan, R.O.C.

ABSTRACT

Efficient and effective response to the requirements of customers is a major performance indicator. Failure to satisfy customer requirements implies operational weaknesses in a company. The traditional method of handling customer requirement for a machine tool manufacturer was dominated by manual process and subjective decision. In this study, we improved the operation process of handling customer requirement. The framework of a customer requirement handling information system for machine tool manufacturers was analyzed, rule-based fuzzy inference was utilized, and a prototype was developed. The system supports both customers and service personnel in providing a systematic way of fulfilling and analyzing customer requirements.
Keywords: Fuzzy inference system, customer requirements, information systems

Keywords

Fuzzy inference system, customer requirements, information systems

1. Introduction

Machine tool enterprises must have reliable operational ability to attract international orders in the highly competitive global environment. The machine tool sector consists of more than 1400 companies in Taiwan, with over 26,000 employees. This sector thus ranks fifth worldwide as far as manufacturing and export are concerned (Perng, Tsai & Lin, 2003). The Taiwanese government thus prioritized this sector as one of ten major industries for intensive development in the first decade of the new millennium. The machine tool sector faces extreme pressure from global competition and information technology change. The major pressure arises from the need for broad delivery of information to everyone who affects business processes with rapid time-to-market and low cost-of-ownership. To confront this challenge, the industry requires business intelligence, not just for a select few, but for everyone - employees, managers, partners, suppliers, customers, and constituents. Increasing demand and hands-on users are rendering the customary model of business intelligence applications, originated within departments and isolated from the enterprise, inefficient and ineffective. Information technologies are anticipated to transform this paradigm, introducing information-rich interactive capabilities to the e-business environment. Furthermore, globalization and information technology affect industries radically. Many companies have become increasingly aware of the significance of managing customer relationships. Concepts such as customer satisfaction, customer-orientation and service differentiation have become critical initiatives in developing the competitive advantages of a company.

However, the findings in Kalakota & Robinson (1999) show, developing a new customer costs more than six times the effort of keeping a present one. Hence, any demand of the existing customer needs to be treated more positively. Customer requirement management seeks to ascertain customer needs and solve or fulfill them. Accelerating response depends on a more advanced information system and the exploitation of a specific method.

In manufacturing industries, cases of utilizing information technology to promote performance are growing significantly. An example of this is the after-sale service management information system for the machine tool industry proposed by Tsai, Shuei, Perng, Lin, and Yao (2001). The system handles data generated from merchandise service, such as repair request and maintenance history. It is a management information system for customer requirement. However, their system is incapable of handling the requirements expressed by customer with vague semantics.

For most situations in machine tool manufactures, the customer requirements are handled by service personnel manually in different department with their subjective recognition and responses. However, this appears tedious when there are diverse customer needs and inconsistent departmental policy. For instance, due to insufficient technical knowledge and lack of awareness about machine specifications, customers may present requests with some inexact meanings. Our interview to major machine tool service managers also revealed that the most significant problems the companies encountered, i.e. customer requirement frequently received with verbally vague description and the customer service is not performed systematically. We noticed and recognized the practical difficulties in handling customer requirement, and thus developed a systematic means of handling them. The detailed analysis was shown in the sections of demand analysis and system analysis. Fuzzy inference was utilized in this study to deal with events with fuzzy demands. When customer requirements are encountered, fuzzy inference is employed to infer the root causes of the problems and suggesting suitable solutions to those problems. Furthermore, this system also helps the customer service department to assess customer feedback, integrating it with after-sale service records. By analyzing the service records, product quality can be improved and the service improved.

Additionally, the future electronic trading environment will center on business networks or community-like business models. Alternatively, information technology enables digitalization of the supply chain. The significance of digitalization resides in synergy resulting from sophistication of external trade environments rather than any immediate benefit from computerization. Conversely, applying digitalization when the environment matures may already be too late (Tsai *et al.*, 2001). Therefore, the motivation to apply information technology to customer requirement management, and thus to develop a customer requirement handling information system is to join the forces of all participants in the entire service process, enhancing prompt customer requirement handling.

This study analyzed and designed a customer requirement handling system for the machine tool industry. The implemented system was thereafter installed on an application server and functioned effectively at a machine tools manufacturer. More specifically, the objectives were as follows: (1) using fuzzy inference to surmount verbal description of requirements from a customer; (2) analyzing information demands of a customer requirement handling system and sequencing the system framework based on fuzzy inference; and (3) constructing a prototype of the customer requirement handling system for machine tool industry.

2. Literature Review

The behavior of consumer requirements has been defined in many studies (Day, 1980; Hart, Heskett & Sasser, 1990; Heskett, Jones, Loveman, Sasser & Schlesinger, 1994; Jacoby & Jaccard, 1981; Joentgen, Mikenina, Weber, Zeugner & Zimmermann, 1999; Lewis, 1982). There are two basic concepts: (1) the behavior of consumer demands is caused by dissatisfaction of the consumers; (2) the response to the behavior of consumer demands divides into two types: behavioral response and non-behavioral response. Behavioral responses include conveying the negative images to his/her friends and complaining directly to retailers, or to the third parties, such as the Consumer Protection Foundation. A non-behavioral response forgets the unsatisfactory experience without taking any further actions. Customer requirements management is a procedure adopted to solve customer requirements and establish customer confidence in a company. The industry uses requirement data to alter product, correct service weaknesses and thus enhance company reputation. Requirement data enables employees to immediately comprehend the customer requirements as problems are encountered. Because customer requirements are handled after the product or service is sold, requirements management is also called service remedy. Gronoos (1988) notes that the tactics of service remedy involve measures adopted when the supplier service fails.

Christo (1997) discovered that the extent of customer satisfaction is inversely proportional to the response time. The longer the customer waiting time is, the more significant the inverse relationship. When customer requirements occur, appropriate and quick requirement management is of importance.

In related studies for the machine tool sector, Tsai, Shuei, Perng, Lin & Yao (2001) and Perng, Tsai & Lin (1996) targeted maintenance service systems. Their studies proposed some related frameworks, but the indistinct features of customer requirements were not discussed. Perng *et al.* (1996) claimed that the difficulty of maintenance service in practice encompasses: (1) training effect not significant, (2) repair process and material unrecorded in maintenance report and (3) component or material for early generation product not managed. Perng *et al.* (1996) also stated that building an information system for machine fault diagnosis raises many difficulties. First, due to the lack of engineering knowledge, maintenance information was not accumulated in practice. Second, the technology revolution leads to a short product cycle, resulting in greater difficulties with repair tasks. The maintenance tasks mainly rely on skilled technicians which take a long period of time to train. Ou (1997) discovered that the major customer complaint of international machine tool buyers related to the lack of quick response to their requests.

Zadeh, who first depicted fuzzy sets in 1965, advocated the application of fuzzy set theory in quantitative measures of the human thinking process (Zadeh, 1975). The practical application system of fuzzy control was developed in Europe. Fang (1995) proposed a method for on-line machine condition monitoring involving a fuzzy feature-state relationship matrices and expert system. Qu and Zhang (1999) stated that fuzzy semantics handling was a widely applied methodology. Chen (2001) used fuzzy set theory to implement a circuitry framework of a rhetoric fuzzy controller.

Various applications of fuzzy inference can be found in the literature (Berkan & Trubatch, 1997; Ruan & Kerre (2000); Chen & Pham, 2000). However, an integrated information system for dealing with customer requirement is not implemented. Liu and Chen (1995) developed a machine troubleshooting expert system through a fuzzy multiattribute decision-making approach. This system consists of five components and improves the efficiency of the diagnostic process. Javadpour and Knapp (2003) focused on the implementing a predictive neural network for use as an operator's aid in diagnosing faults with high prediction accuracy in an automated manufacturing environment. To evaluate the performance of the model, the network was assessed with both simulated time series and real time machine vibration data gathered in lab experiments. However, the fuzzy neural network approach was not applied to customer requirement. Joentgen *et al.* (1999) proposed a clustering-based dynamic method for early recognition of changes in a machine's state and thus for automatic fault detection. This recognition system requires less expert knowledge than traditional approaches. However, the collaborative approach ensuring efficient information sharing was not employed.

3. Demand Analysis

To analyze the functional demand of a customer requirement handling system, and to fully augment system usability, managers from six machine tool manufacturers and sales agents were interviewed. All were managers working in the service departments of manufacturers ranked within the top ten for sales figures from 1999 to 2003 in Taiwan.

3.1 The Machine Tool Sector

The Taiwanese machine tool industry is one of the most potential industrial activities on the island. Five major characteristics of the machine tool industry were isolated from the interview materials.

The machine tool industry has critically close relations with the aerospace industry, electronics, the automobile industry and defense. It plays an important supportive and cooperative role with other industries.

The machine tool industry, producing high precision machining tools, is technology-intensive. Demand for machine tools in market is dramatically influenced by economical cycles. The numbers of components and range of materials is large. Enterprises frequently encounter urgent requests for changes of design. While enterprises need to continuously invest on technology and human resources, the returns on investment are relatively low.

3.2 Essential Elements of the Interview

In-depth interviews were conducted to build up an information system for dealing with customer requirements so that most of the problems mentioned earlier could be solved. Six major companies were selected for analysis. The main analytical contents were as follows: (1) basic information about the interviewed company, (2) the company situation with regard to computer and networking usage, (3) the present process and condition of customer requirement handling and management and (4) the demand for a customer requirement handling system for machine tool service.

The interview conclusion was that the industry urgently needs an information system to handle customer requirement and enhance customer relationships. The interview material also uncovered many key factors and concerns for developing such a system. The most significant problems the companies faced are the customer requirement frequently received with verbally vague description and the customer service are not performed systematically. The operational practice, system analysis and system development in the study will now be discusses in detail.

3.3 Practical Operation Process

The operational process was executed inefficiently in the following manner shown in Figure 1. First, customer requirements were logged via phone, fax or maintenance personnel. Basic customer information such as address, phone number and the machine model was recorded by either the service person in the call center or maintenance personnel. The requirement responses were then analyzed online or sequentially called/faxed back to the customer. The problem-solving hints or instructions were sent to the customer, and the problems were expected to be solved step by step by customers.

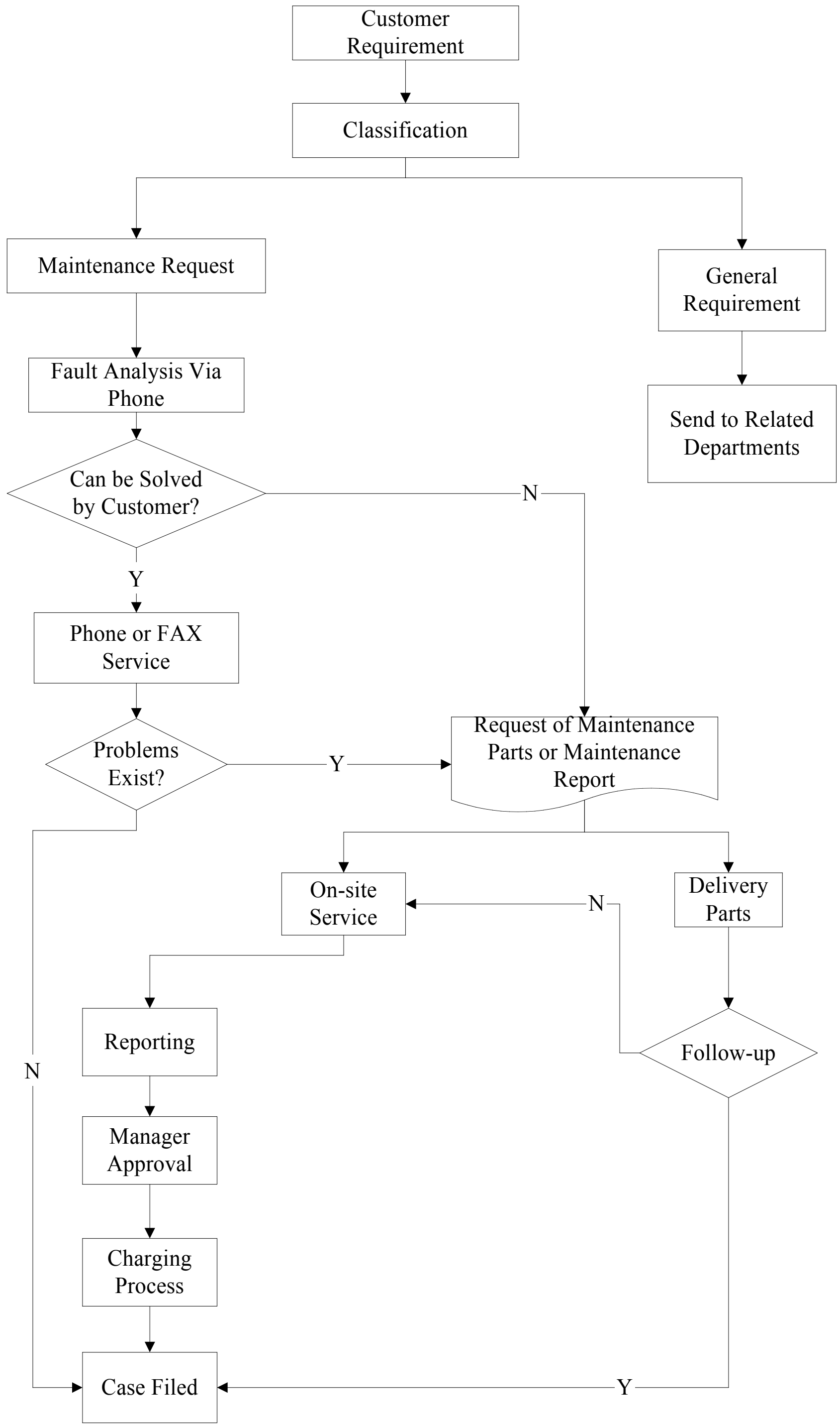


Figure 1. Customer requirement handling process in practice

If the customers could not solve the problem by themselves, the maintenance personnel offered two further forms of service: (1) onsite service and (2) maintenance part delivery service. Before performing onsite service, service personnel needed to fill out a service request form. For those requirements able to be met by delivering parts for self-replacement, the service personnel followed up the situation to ensure the requirement was fulfilled. After service was completed, either onsite or by part delivery, the service personnel reported the process by filling out service records, including service reports, customer response sheets and part request forms. Those sheets or forms were then sent to the department manager for approval. If the requirements were made and fulfilled via phone or fax, the forms and sheets were omitted, that is, no any information was recorded. The process then went to billing, executed by the accounting department, for the service charge, man-hour charge and/or maintenance part charge. The inventory amount was also updated if the parts were checked out from stock.

4. System Analysis and Design

4.1 System Analysis

The industry is not using information systems to assist their service tasks or handle their service information, nor for further analysis of service information. The realization of the need of the industry stimulated the analysis and development of a customer requirement handling system. Developing a customer requirement handling system is largely an attempt to enable the sharing of enterprise resources, such as service information and manpower. With the aid of a customer requirement handling system, the customer service will be completed more efficiently. This will also help release customers from their inconvenient situation, consequently promoting customer royalty and increasing the customer satisfaction. To achieve this goal, an enterprise should provide a standard of operational procedure for maintenance service personnel to work by. Because different manufacturers usually develop very different processes, a standard procedure for customer requirements handling appears to be urgently needed. This investigation addresses the operational process, system analysis and system design for a customer requirement handling system.

4.2 System Flow

Figure 2 illustrates the process of the customer requirement handling system. The system is divided into two subsystems: general requirement handling and machine fault diagnosis. Customer problems first collected and classified. Service personnel directed the problem to the relevant subsystem triggering the handling process. For general requirements, the classification process is selected, fuzzy semantic treatment is instituted if needed, and the problem is recorded in the requirement recording system.

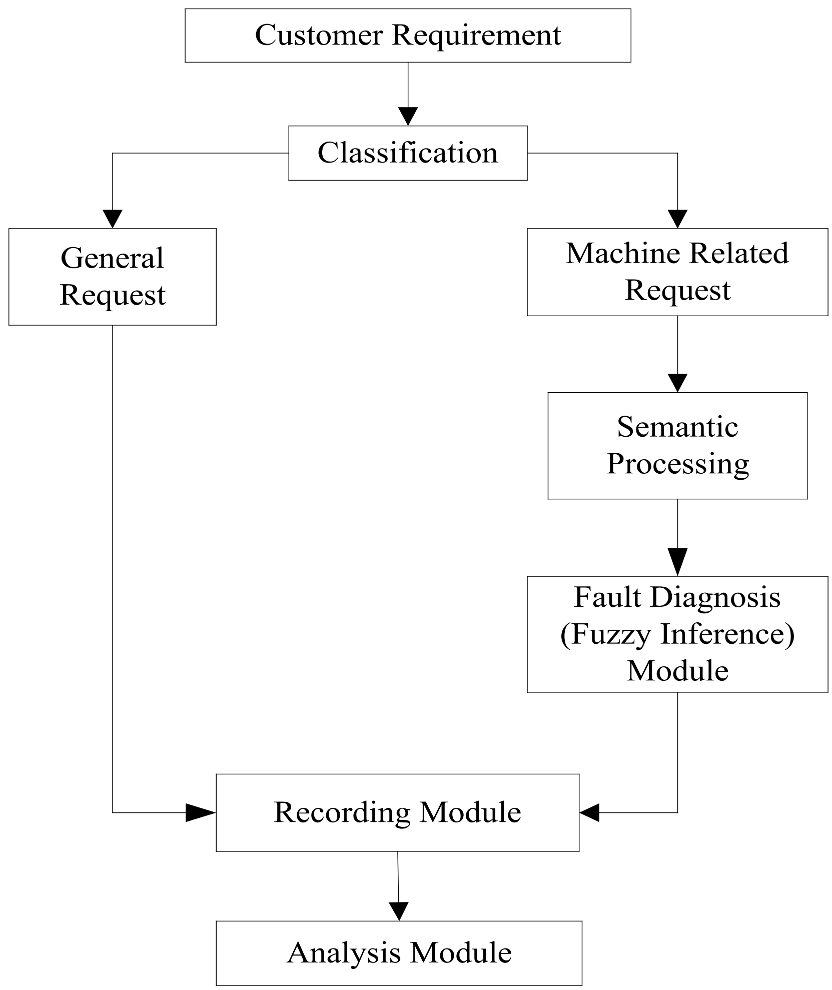


Figure 2. System flow

For machine fault related requirements, the machine fault diagnosis system will be activated. The verbal expression from a customer is analyzed and mapped to a system terminology. Then, the fuzzy inference process will be activated.

Following processing the general requirement or machine fault diagnosis, system stores the requirement treatment details in the requirement recording system. Finally, the data in the requirement recording system are conveyed to the requirement analysis system for further evaluation. Some statistical computing was implemented in the prototype system, such as the defect rate, or requirement rate of a certain customer in a particular period of time. Statistical computing can be extended to many aspects, and is supportive of decision making.

Based on the proposed operational process and the functional demand of a customer requirement handling system, a system analysis was conducted with IDEF0, which is a structured system analysis tool. In the IDEF0 model, activities can be described by their inputs, outputs, controls, and mechanisms. In addition, the description of activities can be recursively refined into greater and greater detail until the model is as descriptive as necessary for implementation (Bravoco & Yadav, 1985; Mayor, Benjamin, Bruce & Painter, 1995; Perng, Tsai & Lin, 2003; Ross, 1985).

A system analysis result was derived for efficient, economical and customized establishment of a customer requirement handling system. The result was further extended to the system design stage, as to constructing a prototype illustrating the feasibility of the framework. Figures 3 and 4 are two examples indicating the system processes depicted by IDEF0. The system contains four sub systems, namely requirement classification system (RCS), fault diagnosis system (FDS), requirement recording system (RRS), and decision analysis system (DAS).

A0 Customer Requirement Handling System
A1 Requirement Classification System
- A11 Human Related Complaint
- A12 Machine Related Complaint

A2 Fault Diagnosis System
- A21 Fuzzy Inference System
- A22 Expert System
- A23 Rule base Management System

A3 Requirement Recording System
- A31 Customer Tracking System
- A32 Personnel Recording System
- A33 Data Query System

A4 Requirement Analysis System
- A41 Breakdown Rating Analysis
- A42 Maintenance Supplies Analysis
- A43 Machine Performance Rating
- A44 Personnel Performance Rating

Figure 3. System framework in IDEF0 structure

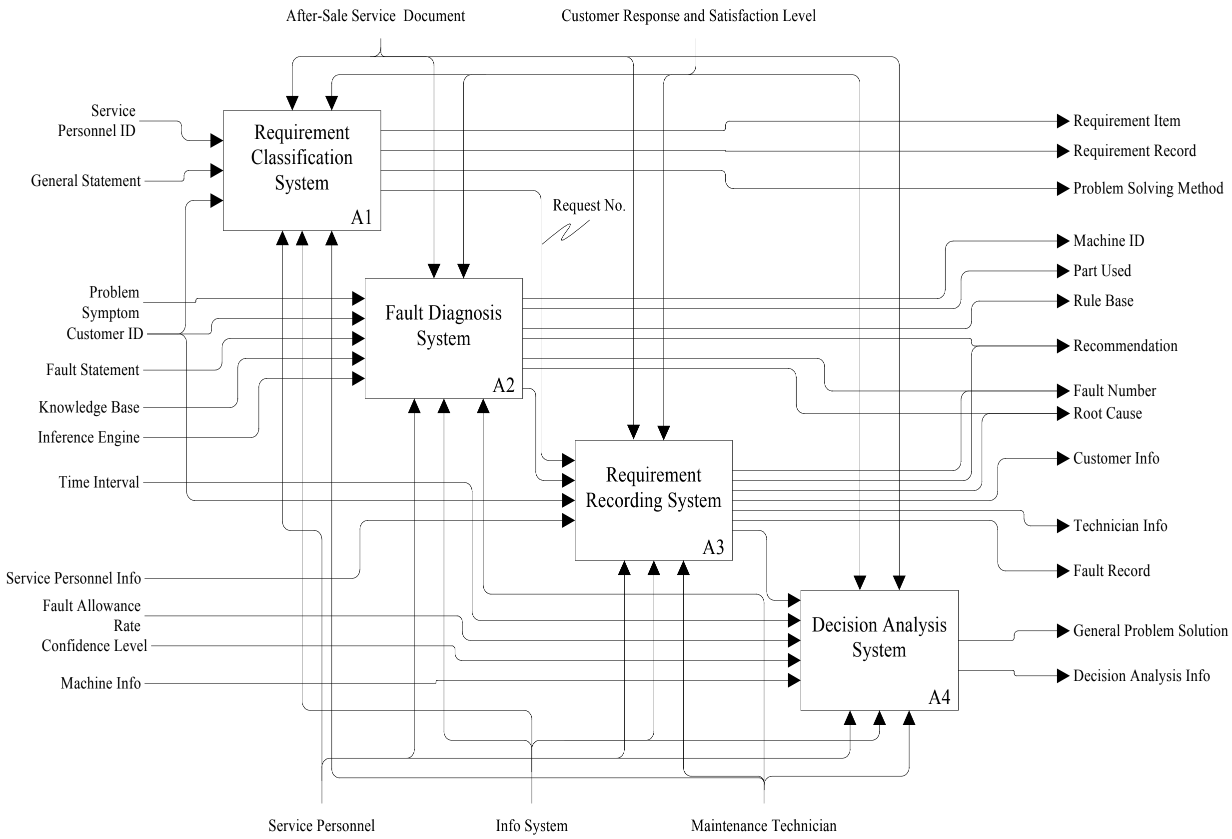


Figure 4. System process depicted by IDEF0

4.3 Fuzzy Inference

This system incorporates rule-based fuzzy inference. Fuzzy inference in this system refers to the use of computer programs to execute inference work resembling what humans do daily. Inputs and outputs are two basic elements in a system using fuzzy handling approaches. The input constitutes some ambiguous verbal semantics or unclear concepts for a specific event, such as the vague description of motor temperature or the leakage level of an oil pump. Following the fuzzy inference mechanism, the output can be a fuzzy set or a precise set of certain features. Fuzzy inference infers the results from the existing knowledge base. For after-sale service personnel, it is an important step to know about the customer requirement thoroughly when solving requirement problems. They usually spend considerable time finding out the truth of a customer's requirement, or the original customer intention. To conquer this problem, a method of synthetic fuzzy evaluation was adopted to construct the inference mechanism in this study. Figure 5 illustrates the framework of the fuzzy knowledge base and the relationship between elements. The fuzzy knowledge base comprises fuzzy concept base, fuzzy proposition base, fuzzy rule base and fuzzy strategy base.

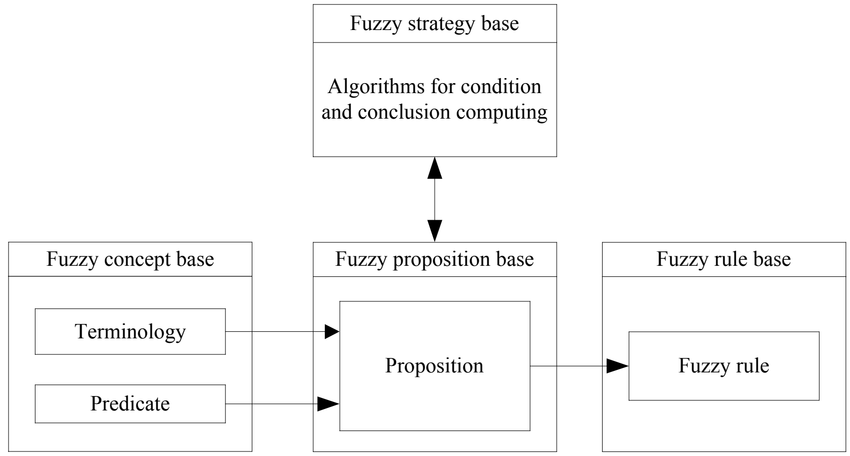


Figure 5. Framework of Fuzzy Knowledge Base

Fuzzy concept base. This contains the terminology and relevant predicate of a verbal expression. Terminology is in the domain of the fuzzy set, possesses many pre-defined dismemberment values denoted by predicates.

Fuzzy proposition base. Membership functions accrue to the fuzzy proposition, which was induced from fuzzy concept base. There are numerous types of membership functions, such as S-shape, Z-shape, and Π-shape, all easily definable with equations and parameters. For example, if the general fuzzy set is expressed as

$$A = \{(x, \mu_A(x))\}, \quad x \in X$$

where μ denotes the membership function, and $(x, \mu_A(x))$ is a singleton, then a fuzzifier given by

$$\mu(x) = \frac{1}{1 + (x / K_2)^{-K_1}}, \quad x \in X$$

produces an S-shaped curvature. K_1 and K_2 are called the exponential and denominational fuzzifiers, respectively. By having controllable parameters such as K_1 and K_2, adaptive fuzzy algorithms can be developed.

Fuzzy rule base

The fuzzy proposition is then presented in IF-THEN format and constitutes the rule base. Specifically, a finite fuzzy logic implication statement in the rule base was described by a set of general fuzzy IF-THEN rules containing only the fuzzy logical AND operation, in the form "IF a_{11} *is* A_{11} AND …AND a_{1n} *is* A_{1n} THEN b_1 *is* B_1."

Fuzzy strategy base. This contains the algorithms for computing the condition part and the conclusion part. A proposition might encompass many conditions. An appropriate fitness of a rule had to be found so that the conclusion can be drawn. This is carried out by a process of implication. A membership function that defines the implication relation can be expressed in a number of ways. To illustrate the operation, we assume that we have the following simple conditional proposition (canonical rule):

IF X is A THEN Y is B

The implication relation is defined by

$$R(x,y) = \int_{x,y} \mu(x,y)/(x,y)$$

where linguistic/fuzzy variable X and Y take the value of A and B, respectively, and $\mu(x,y)$ is the membership function of the implication relation. The membership function is denoted by

$$\mu(x,y) = \mu_A(x) \wedge \mu_B(y)$$

The symbol $\wedge$ correspond to intersection operation.

4.4 Numerical Illustration

To demonstrate the operation of these elements, an example is given as follows. The fuzzy concept base has a terminology of "oil leakage" with five terms of predicates, namely very few, few, fair, much, very much. The membership function for each predicate level is provided in Figure 6. In this example, "very few" is an X-shape function, "very much" an S-shape, and Π-shape for the others.

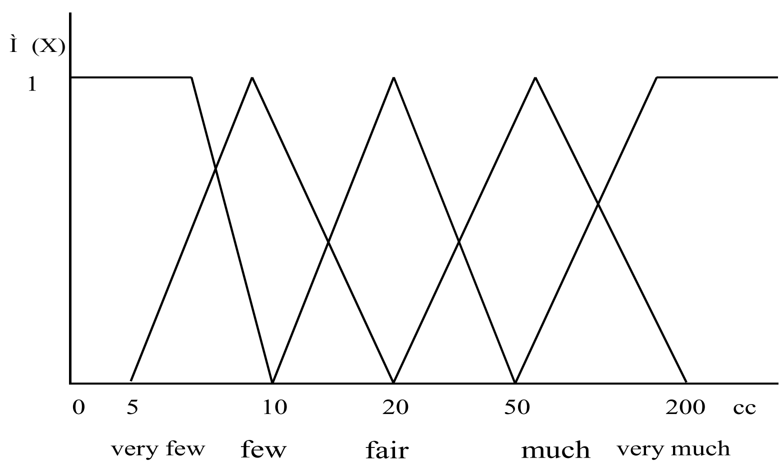


Figure 6. Membership functions for oil leakage

The requirements were received from a customer and stated as "We have a problem of the oil pump. It leaks seriously and is not working. The temperature of the pump is as high as 280°F. The leakage is about 60 c.c." The system will fuzzify the input by searching the fuzzy proposition base and determining two predicates as following:

1. Pump temperature, x denoted in Fahrenheit, is *high*:

$$T = \begin{cases} \dfrac{x-200}{150}, & 200 < x \le 350 \\ \dfrac{500-x}{150}, & 350 < x \le 500 \end{cases}$$

2. Volume of oil leakage, y denoted in c.c., is *much*:

$$L = \begin{cases} \frac{y-20}{30}, & 20 < x \leq 50 \\ \frac{200-y}{150}, & 50 < x \leq 200 \end{cases}$$

The system then searched the fuzzy rule base to retrieving related rules. Two rules were found as follows:

1. Rule 72: IF pump temperature is high AND oil leakage is much THEN oil pump is broken.
2. Rule 163: IF oil leakage is much AND motor noise is high THEN parking ring is damaged OR pipe screw is loose.

Rule 72 contains two predicates of the requirement statement, *much* and *high*. The implication computation of H and L was then performed in the fuzzy strategy base*:*

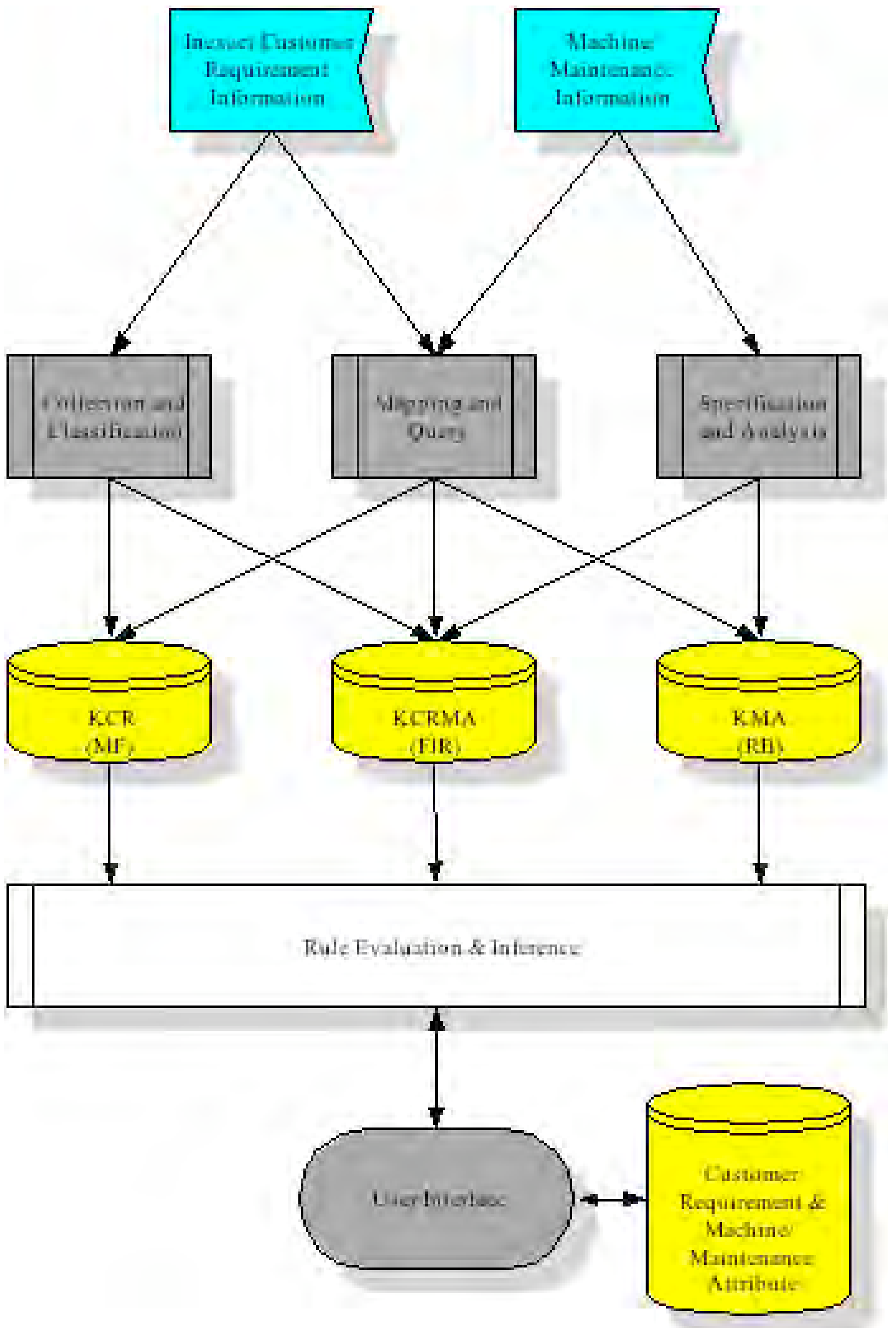
$T \times L = R(x, y) = \int \mu_T(x) \wedge \mu_L(y)/(x, y)$,

yielding product of 0.86667.

Rule 163 contains one predicate of the requirements statement. The computational result of membership is 0. The THEN operator performs a mapping-like function. After comparing those two rules, rule 72 was adopted to infer the root cause of the problem and achieve the result of "oil pump is broken."

4.5 System Architecture

The system performs many activities to handle customer requirement, such as information collection, classification, fuzzification, rule evaluation, inference, aggregation and defuzzification as depicted in Figure 7. The processes result in crisp target values for handling customer requirement or root causes of machine faults. When requirement information has been collected and the customer information gathered, the system activates the classification module to group the request. When a set of customer requirements has been identified as a machine-related requirement, the attributes are first treated according to the term sets (sets of linguistic variables) of the relevant terminology and transformed by the fuzzy set hedges as appropriate to fully elaborate the maintenance attributes, as has been depicted in the previous section in this study. This process was applied to all the ambiguously described customer requirements. Using the inferred fact, an intact fuzzy rule base was developed, to capture the customer requirement knowledge and experience, defining the relationship between customer requirements and machine diagnosis characteristics. The rule base for machine fault diagnosis was constructed by domain experts. The system is capable of processing customer requirements and providing proper recommendations to the specific requirement.



KCR: Knowledge about Customer Requirement
KMA: Knowledge about Machine Attributes
KCRMA: Knowledge about the relationship between Customer Requirement and Machine Attributes
MF: Membership Functions
RB: System Rule Base
FIR: Fuzzy Inference Rules/Propositions

Figure 7. Architecture of fuzzy inference system

System Implementation

The prototype system was developed by many software packages: Access, Visual Basic, and Active Server Pages (ASP). Access helped design the database. Visual Basic was used to construct the user interface, while ASP facilitated connection and modification of the system database, integrating the whole system. The prototype system is running on a web server in a local machine tool manufacturer. After test activity and system refinement, the performance of the system is promising.

4.5.1 System Operation Configuration

The system is configured for installation at the manufacturing site. It is installed in a web server running database applications. The system users are primarily customers and maintenance personnel of the machine tool manufacturer. The system database is updated while the system receives commands from remote sites, such as sales agents, maintenance centers or authorized end customers. It also responds to the managers of the manufacturers while they perform inquiries such as machine performance analysis or statistical analysis of customer requirement.

4.5.2 Design of Database

The design of database for the prototype system is mainly based on relational database with Access. The setting of primary keys and database normalization are inevitable for implementing a database application system successfully.

The primary key is a field that uniquely describes each record. In the Requirement Classification System, for instance, the customer ID number is set to be the primary key in the data table storing customer information. The content of this field is unique and generated by the system automatically. There are seven major tables used in the system database, including customer table, requirement table, personnel table, components table, fuzzy rule base related tables, and knowledgebase related tables. Primary keys are associated with a field in other tables according to a specific relation.

4.5.3 Design of Interface

The system interface was designed on the basis of system usability. It allows users to input data by clicking on a tag and making their choice directly from the list. Figure 8 presents a display image for fuzzy handling of a customer requirement.

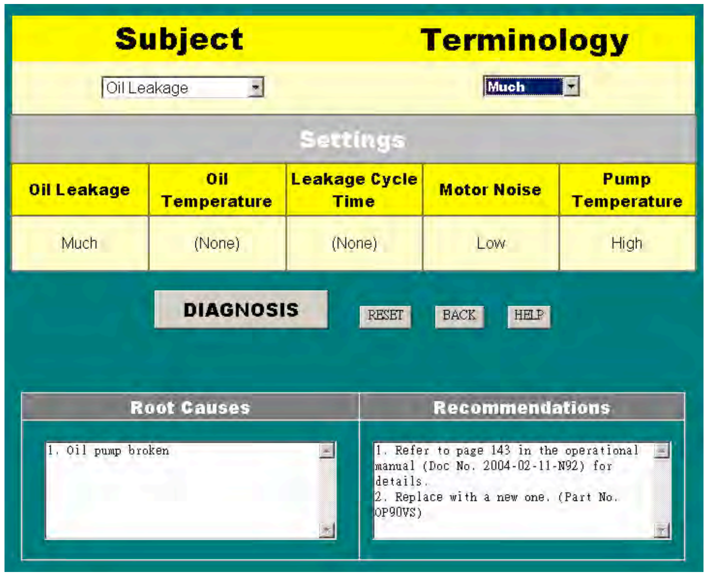


Figure 8. Design of system interface

5. Conclusion

Requirement handling activities play an important role in after-sale service within the machine tool industry. After-sale service is a critical factor in maintaining customer royalty. During recent years, the development of information systems has assisted the companies in solving many managerial problems, to keep business running efficiently. It is useful and valuable to use information systems to handle customer requirement problems. This study proposes and develops an information system for customer requirement handling in the machine tool industry. Fuzzy inference is utilized in this study. When a customer requirement is encountered, fuzzy inference is adopted to articulate the unclear parts of requirement wordings, to infer the root causes of problems, and to suggest an appropriate solution to those problems.

Furthermore, this system also helps the customer service department to analyze customer feedback and integrate it with after-sale service records. By analyzing the service records, product quality can be improved and the service can be enhanced.

Recent advances in information technology, companies have provided an opportunity for significant improvement in customer requirement handling. However, several important issues still need to be adequately addressed by future work. The system developed in this study also needs more refinement. Further work will direct to employing expert system and enhancing inference capacity of the expert system, as well as performing more functional analysis from the database.

References

Bally, L., Brittan, J. & Wagner, J. H. (1997). A Prototype Approach to Informational System Design and Development, Information and Management, 1, 21-26.

Berkan, R.C & Trubatch, S.L. (1997). Fuzzy Systems Design Principles - Building Fuzzy If-Then Rules Bases, NJ: IEEE Presss.

Bravoco, R. & Yadav, S. (1985). Methodology to Model Functional Structure of an Organization, Computers in Industry, 6, 345-361.

Chen, C.-Y. (2001). Framework and Circuitry Design for a Fuzzy Logic Controller, Unpublished Master Thesis, National Cheng-Kung University.

Chen, G. & Pham, T.T. (2000). Introduction to Fuzzy Sets, Fuzzy Logic, and Fuzzy Control Systems, FL: CRC press.

Chi, Y.-C. & Zhang, X.-D. (1999). Fuzzy set theory and Applications, Taipei: Ruling Publishing Co.

Cristo, B. (1997). An Experimental Study of Service Recovery Options, International Journal of Service Industry Management, 8, 110-130.

Davis, F.D. (1993). User Acceptance of Information Technology: System Characteristics, User Perceptions and Behavioral Impacts, International Journal of Man-Machine, 475-487.

Day, R.L. (1980). Research Perspectives on Consumer Complaining Behavior, Theoretical Development in Marketing, 211-215.

Fang, X.D. (1995). Expert System-Supported Fuzzy Diagnosis of Finish-Turning Process States, International Journal of Machine Tools and Manufacture, 35(6), 913-924.

Gronroos, C. (1988). Service Quality: The Six Criteria of Good Perceived Service Quality, Review of Business, 9, 10-13.

Hart, W.L., Heskett, J.L., & Sasser, W.E. (1990). The Profitable Art of Service Recovery, Harvard Business Review, 68,148-156.

Heskett, J.L., Jones, T.O., Loveman, G.W., Sasser, W.E., & Schlesinger, L.A. (1994). Putting the Service-Profit Chain to Work, Harvard Business Review, 72, 164-174.

Jacoby, J. & Jaccard, J.J. (1981). The Source, Meaning and Validity of Consumer Complaining Behavior: A Psychological Review, Journal of Retailing, 57, 4-24.

Javadpour, R. & Knapp, G.M. (2003). A Fuzzy Neural Network Approach to Machine Condition Monitoring, Computers & Industrial Engineering, 45, 323-330.

Joentgen, A., Mikenina, L., Weber, R., Zeugner, A., & Zimmermann, H.J. (1999). Automatic Fault Detection in Gearboxes by Dynamic Fuzzy Data Analysis, Fuzzy Sets and Systems, 105, 123-132.

Kalakota, R. & Robinson, M. (1999). E-Business Roadmap for Success, NJ: Addison-Wesley.

Landon, E.L. Jr. (1980). The Direction of Consumer Requirements Research, Advances in Consumer Research, 7, 335-338.

Lewis, R.C. (1982). Consumers Complaint: What Happens When Business Responds? Proceedings of Consumer Satisfaction/Dissatisfaction and Complaining Behavior Conference, 88-94.

Liu, S.-Y. & Chen, J.-G. (1995). Development of a Machine Troubleshooting Expert System via Fuzzy Multi-attribute Decision-Making Approach, Expert Systems with Applications, 8(1), 187-201.

Maute, M.F. & Forrester, W.R. Jr. (1993). The Structure and Determinants of Consumer Complaint Intentions and Behavior, Journal of Economic Psychology, 14(2), 219-247.

Mayor, R.J., Benjamin, P.C., Bruce, E. & Painter, C.M. (1995). A Framework and a Suit of Methods for Business Process Reengineering, Business Process Change: Reengineering Concept, Method and Technologies, Grover, V. & W. J. Kettinger Ed., Idea Group Publishing.

Ou, J. (1997). The perspectives of machining tool industry in Taiwan, Special Issue for Machining Tool Industry, 9th Ed., Taipei: Economy Daily.

Perng, C., Tsai, J.-T., & Lin, S.-S. (2003). A Design of Collaborative Commerce Service System: Taiwan Machine Tool Industries as an Example, Proceedings of the Portland International Conference on Management of Engineering & Technology (PICMET), Portland, OR, USA.

Pressman, R.S. (1992). Software Engineering: A Practitioner's Approach, 3rd ed., NY: McGraw-Hill Inc.

Ross, D.T. (1985). Application of SADT, IEEE Transactions on Computers, April, 25-34

Ruan, D. & Kerre, E.E. (2000). Fuzzy If-Then Rules in Computational Intelligence - Theory and Applications, MA: Kluwer Academic Publishers.

Sih, P.-S. (1994). Guide to Object-Oriented Technology, Taipei: Infocomp Publishing Co.

Tax, S.S. & Brown, S.W. (1988). Recovering and Learning from Service Failure, Sloan Management Review, 75-88.

Tsai, J.-T., J.-C. Shuei, C. Perng, S.-S. Lin, and M.-C. Yao (2001). An Information System of Maintenance Component Management for Machine Tool Industry, Proceedings of the 2001 Annual Conference for Chinese Institute of Industrial Engineer.

Tsai, J.-T., C.-C. Lu, C. Perng, S.-S. Lin, and H.-C. Chu (2001). A Structure of the Marketing Executive Information System for the Machine Tool Industry. Proceedings of the 2001 Annual Conference of Chinese Institute of Industrial Engineers.

Zadeh, L.A. (1975). The Concept of a Linguistic Variable and Its Applications to Approximate Reasoning, Part 1, 2 and 3, Information Science, 8:199-249, 8:301-357, 9:43-80.

Knowledge Management: Motivating Strategic Behavior

Tracy A. Hurley, Texas A&M University-Kingsville, San Antonio, USA
Carolyn W. Green, Texas A&M University-Kingsville, San Antonio, USA

Abstract

According to the resource-based view of the firm, knowledge can be a source of competitive advantage for many organizations. This study explores the role of both intrinsic and extrinsic rewards on knowledge creation and knowledge sharing behaviors. A sample of 137 employees reveals that intrinsic motivational factors positively affect knowledge creation behavior. This supports the assertions presented by both Maslow's (where knowledge workers are motivated by higher-order needs) and Herzberg's (where knowledge workers are motivated by challenging work) theories. Social Exchange theory propositions are also supported with respect to the positive effects of extrinsic motivational factors on knowledge sharing behavior. Through the careful combination of both intrinsic and extrinsic motivational tools, managers can develop a culture that encourages knowledge management activities. These activities (or strategic behaviors) can provide an organization with a competitive advantage and the means to maintain this advantage through knowledgeable workers.

Keywords

knowledge management, knowledge sharing, knowledge creation, motivation, resource-based view

Knowledge Management (KM) activities such as knowledge creation and sharing have received much attention in a variety of organizational disciplines such as human resources/organizational behavior (e.g., culture, motivation, rewards), organizational theory (task and organizational structure), information technology (e.g., systems, software), and strategy (competitive advantage). Essentially, the studies approach these concepts from a variety of directions but examine similar constructs that lead to the identification and examination of "crucial variables in the creation, dissemination and implementation of new customer-oriented knowledge" (Selen, 2000, p.353). The strategy literature is instrumental in that it provides a justification for the development of a stream of research for KM. The Resource-Based View

(RBV) of the firm examines KM in terms of its importance in securing and maintaining a competitive advantage. Organizational behavior, human resources, and sociological theories address the employee "care-why" through the examination of motivation and rewards. Davenport and Prusak (1998) suggest that high levels of "care-why" significantly enhance the overall performance of the firm. If knowledge is viewed as a competitive advantage then it is important to determine how to motivate employees to engage in this strategic behavior. This paper examines knowledge management behavior in terms of knowledge creation and sharing using both RBV and motivation theories. Specifically, the impact of intrinsic and extrinsic motivational factors are examined to determine whether they motivate the KM (i.e., strategic) behaviors of knowledge creation and knowledge sharing.

Resource-Based View Theory

RBV theory of the firm suggests that knowledge is a valuable resource and a source of competitive advantage (e.g., Barney, Wright & Ketchen, 2001; Wright, Dunford & Snell, 2001; Bollinger & Smith, 2001; Meso & Smith, 2000; Selen, 2000). RBV theory proposes that resources must be rare, valuable, inimitable, and non-substitutable in order to be a source of sustainable competitive advantage (Barney, 2001). Based on this foundation, Meso and Smith (2000) argue that, "KM can be viewed as the creation of sustainable competitive advantage through continued organizational learning" (p. 226) and Conner and Prahalad (1996) argue that, "privately held knowledge is a basic source of advantage in competition" (p. 477).

Coupled with the idea of knowledge as a source of competitive advantage, it is important to understand how and why this knowledge is shared among employees. The culture of an organization plays a central role in this process. Meso and Smith (2000) define culture as "the shared beliefs, norms, ethics and practices within an organization" (p.232). They further define a "knowledge-friendly culture" as one in which "employees highly value learning and exhibit a positive orientation to knowledge" and argue that culture is "intangible" and "unique" and "cannot be replicated, imitated, acquired, or substituted" (p. 232). This suggests that organizational culture is also a source of competitive advantage. Since organizational culture and knowledge can both be viewed as sources of a competitive advantage, RBV suggests that a critical component in an organization's success is the creation and development of processes that sustain a KM culture -- a culture that motivates employees to create and share knowledge. Through the examination of how organizations can use both intrinsic and extrinsic motivational tools to create a KM-friendly culture, company managers can better understand how to motivate strategic behavior (i.e., knowledge creation and sharing). This knowledge can be used to help them build a competitive organization as well as a more knowledgeable workforce thereby increasing their ability to remain competitive in the future.

Intrinsic Motivation and Knowledge Creation

Motivation theory has been reviewed several times within a KM environment. Several different models have been used to examine the expected outcomes predicted by theory. For example, Maslow's need hierarchy has been examined with respect to motivating knowledge workers. Stott and Walker (1995) as well as Tampoe (1996) suggest that knowledge workers are motivated by the highest three hierarchical levels – social needs (affiliation, love), ego needs (status, respect of peers), and self-fulfillment needs (self-actualization). In addition, Wilkinson, Orth, and Benfari (1986) suggest that in highly professional organizations, such as R&D groups, many workers are motivated by these same three higher-order needs. Hendricks (1999) suggests that this implies that knowledge workers are motivated by their desire for self-actualization.

Herzberg's two factor theory proposes that when a "job presents the worker with opportunities to satisfy higher needs, he will be motivated to better work performance" and "management should provide jobs that offer opportunities for achievement, recognition, challenge, growth, responsibility, and accomplishment" (Wilkinson, et al, 1986, p. 28). This suggests that if the job does not present intrinsic motivational opportunities such as challenging work or recognition, workers will not be motivated to create knowledge. These sources of motivation are consistent with Maslow's higher-order needs.

Herzberg's model also proposes that there are "hygiene" factors such as salary, status, company policy, and interpersonal relations. Hygiene factors are not motivators but their absence will lead to job dissatisfaction. Maslow's hierarchy and Herzberg's two factor theory, suggest that knowledge creation is motivated by opportunities for personal growth – intrinsic factors (Herzberg's motivators and Maslow's higher-order needs) – and not by financial rewards – extrinsic factors (Herzberg's hygiene factors and Maslow's lower-order needs). Hendriks (1999) suggests that "the challenge of work triggers people to create knowledge" (p. 98).

Hypothesis 1

H_0: Intrinsic motivation has no effect on knowledge creation activities.

H_A: Intrinsic motivation has a positive effect on knowledge creation activities.

Extrinsic Motivation and Knowledge Sharing

Although sometimes combined together as one generic knowledge management behavior, knowledge creation and knowledge sharing are two different KM activities (Alavi & Leidner, 2001). While Hendriks (1999) and Herzberg's two-factor theory suggest that pay will not motivate knowledge workers, social exchange theory suggests that pay can be an important factor in determining an employee's behavior (e.g., knowledge sharing) (Gardner, Van Dyne, & Pierce, 2004). Social exchange theory assumes that 1) actors provide outcomes to each other through exchange, 2) actors are motivated to obtain more of the outcomes that they value and others control, and 3) exchanges between actors recur over time (Molm, Takahashi, & Peterson, 2003). Cropanzano and Greenberg (1997) suggest that when one exchange partner receives a high level of benefits (e.g., pay incentives) from another partner (e.g., the organization), social exchange theory would predict that a reciprocal response (such as knowledge sharing) would occur as long as both partners perceive that the exchange is fair. This suggests that knowledge workers are likely to engage in knowledge sharing activities if they are compensated for their efforts and the relationship (e.g., the reward system) continues over time. Gardner et al. (2004) support this positive relationship between rewards and performance.

Hypothesis 2

H_0: Extrinsic motivation has no effect on knowledge sharing activities

H_A: Extrinsic motivation has a positive effect on knowledge sharing activities.

Method

The data used in this study were collected via an online survey sent to approximately 1500 employees in a variety of non-governmental, nonprofit organizations. These employees were sent an email request by a regional nonprofit funding center asking managers and professional staff to complete the survey. In total, 152 employees answered the survey partially, with a complete set of responses provided by 137. The items included represent measures of intrinsic motivation factors (e.g., reputation improvement, satisfaction, and career advancement) and extrinsic motivation factors (e.g., pay and job security) as well as knowledge creation and knowledge sharing behaviors (see Table 1).

Respondents represented a variety of organizations; they also varied in age and gender. The majority of respondents were over 40 years old (80%, $n = 88$). Seventy-six percent of the respondents were female ($n = 94$) and 24% were male ($n = 30$). Approximately 48% of respondents worked in an organization with 50 or fewer employees ($n = 59$); 24% worked in organizations that had between 51 and 100 employees ($n = 30$); 28% worked in organizations that employed more than 100 employees ($n = 35$). Forty-seven percent had worked for their present employer for one to four years ($n = 58$), while 41% had worked with their current employer for five years or more ($n = 51$). The operating budgets for these organizations ranged from less than \$100,000 ($n = 9$; 7.3%), between \$100,000 and \$500,000 ($n = 26$; 31%), between \$500,000 and \$5 million (n = 47; 39%) and more than \$5 million accounted for 23% ($n = 28$) of the organizations. All respondents were either managerial or professional staff employees.

Results

The means, standard deviations, and correlations of all variables included in the study are presented in Table 2. Intrinsic and extrinsic rewards were measured on a 7-point scale. Intrinsic motivational factors were the highest among these knowledge workers (μ= 6.31, σ=.76 for intrinsic motivational factors for knowledge creation and μ=6.21, σ=.82 for intrinsic motivational factors for knowledge sharing). Similar amounts of both knowledge creation (μ=3.76, σ=.57) and sharing (μ=3.60, σ=.72) behavior were reported. These behaviors were measured on a 5-point scale. Scale reliabilities are presented in Table 1.

Scale reliabilities (Cronbach's alphas) are somewhat low -- most probably due to the low number of items in the scales (Peterson, 1994). To address this issue, the analysis is broken into three sections. The first section examines the hypotheses as predicted with the scales as developed. The second section combines both scales for intrinsic motivational factors (for both knowledge creation and knowledge sharing) as well as both scales for extrinsic motivational factors to provide one scale for intrinsic motivation (IM) and one scale for extrinsic motivation (EM). Although extrinsic motivation scales were not particularly low, combining them provides consistency with the intrinsic scales. By combining these scales, the reliabilities were improved. The hypotheses were then re-tested in an attempt to validate the original hypotheses. The third section addresses the dependent variables and their lack of high reliabilities. In this section, the dependent variables are combined to form a single variable for knowledge management (KM). The hypothesized relationships are examined using the newly created scales for all variables. This discussion follows in more detail later.

Data Analysis – Section One

Multiple regression analyses were used to test the expected relationships between the variables. The results are summarized in Tables 3 and 4. The first model tested the effects of both intrinsic and extrinsic motivational factors on knowledge creation behaviors. The result was a significant model (F (2, 135) = 5.2, $p < .01$) explaining 6% of the variance in knowledge creation behaviors. Examination of the separate effects of intrinsic motivational factors upon knowledge creation behaviors revealed a positive relationship (*t* =2.92, p < .01) and no relationship was found between extrinsic motivational factors and knowledge creation behaviors (*t* =.31, ns).

The second model tested the effects extrinsic motivational factors on knowledge sharing behaviors. The result was a significant model (*F* (2, 135) = 8.93, p < .001) explaining 10% of the variance. Examination of the separate effects of extrinsic motivational factors upon knowledge sharing behaviors revealed a positive relationship (*t* =3.53, p < .001) while no relationship was found between intrinsic motivational factors and knowledge sharing behavior (*t* =1.04, ns).

Data Analysis – Section Two

As was mentioned previously, the scales for intrinsic motivational factors were low – most probably due to the low number of items in each scale (Peterson, 1994). When the two scales are combined, the reliability for intrinsic motivation (Cronbach's alpha) improves significantly (α = .75). Upon closer examination of the correlations between the original scales used to measure intrinsic motivational factors for both knowledge creation and knowledge sharing and the new intrinsic motivation (IM) scale are high (*r* =.940, *r* =.947, respectively) (see Table 5). This provides some support that the original scales are more valid than their original reliabilities would suggest.

Although the reliability of the scales for extrinsic motivational factors were not particularly low, when these two scales (for both knowledge creation and knowledge sharing) are combined, the reliability for extrinsic motivation improves (α=.89). Combining these scales also provides consistency with the intrinsic motivation scales. An examination of the correlations between the new extrinsic motivation (EM) scale and its original counterparts for both knowledge creation and knowledge sharing also reveals high correlations (*r* = .971, *r* = .970, respectively) (see Table 2).

Using these two new scales for intrinsic motivation and extrinsic motivation, two new multiple regression analysis were performed. The first regression equation examined the first hypothesis that knowledge creation behavior is motivated by intrinsic motivational factors. This hypothesis was supported (*F* (2, 135) = 4.97, p < .01) explaining 5.5% of the variance (See Table 5). Examination of the separate effects of intrinsic motivational factors upon knowledge creation behaviors revealed a positive relationship (*t* = 2.56, p < .05) while no relationship was found between extrinsic motivational factors and knowledge creation behavior (*t* =.77, ns).

The second regression equation examined the second hypothesis that knowledge sharing behavior is motivated by extrinsic motivational factors. This hypothesis was also supported (*F* (2, 135) = 9.12, p < .001) explaining 10.6% of the variance (see Table 6). Examination of the separate effects of extrinsic motivational factors upon knowledge sharing behaviors revealed a positive relationship (*t* =3.02, p < .01) while no relationship was found between extrinsic motivational factors and knowledge sharing behavior (*t* = 1.72, ns).

Data Analysis – Section Three

Although suggested as acceptable in preliminary research by Nunnally (1967), reliabilities of less than .7 seem low for the scales measuring knowledge creation (KC) and knowledge sharing (KS). The lower values of these reliabilities may be due to the relatively small number of items included in the scales. As in section two of this data analysis, the scales for knowledge creation and knowledge sharing behavior were combined to create a new variable --knowledge management behavior (KM). As in section two, this improved the reliability for this dependent variable to .73. An examination of the correlations between the original scales for knowledge creation and knowledge sharing behavior with the new scale for knowledge management behavior indicates high correlations (r = .877 and r = .856, respectively) – again, providing some support for the validity of these scales beyond what is suggested by the original scale reliabilities.

One multiple regression analysis was performed (because the dependent variables were combined) to test the hypothesis that knowledge management behavior is motivated by both intrinsic and extrinsic factors. This hypothesis was supported (F (2, 135) = 8.57, $p < .001$) explaining 10% of the variance (See Table 7). Examination of the separate effects revealed positive relationships between both intrinsic (t = 2.52, $p < .05$) and extrinsic motivational factors (t = 2.14, $p < .05$) upon knowledge management behavior.

Discussion

Knowledge is a source of competitive advantage and as noted above, this concept is supported by several researchers using RBV theory. Knowledge can be a rare, valuable, inimitable, and non-substitutable organizational resource – but only if it does not belong to just one individual within an organization. If knowledge is held closely by one or a few individuals, when they leave the organization the knowledge leaves as well. To maintain a competitive advantage, it is important to ensure that this knowledge is shared with many workers within the organization. Maintaining a competitive advantage can be accomplished through knowledge sharing, but true advances in innovation and creativity that lead to future competitive advantages come from knowledge creation. Knowledge creation involves the development of new ideas through a thoughtful examination of relevant information, concepts, and theories and an exchange of knowledge with other knowledge workers.

Because of the importance of securing and maintaining a competitive advantage, it is necessary for organizations to determine how to best motivate strategic behavior among employees. The results of this study support the hypothesis that knowledge workers are motivated to create new knowledge because of intrinsic motivational factors such as satisfaction and reputation. These results support Herzberg's contention that pay is not a motivator and that workers are motivated by intrinsic factors. The results also support the assertion that knowledge workers are motivated by Maslow's higher-order needs as suggested by Stott and Walker (1995), Tampoe (1996), and Wilkinson, et al, (1986). The analysis reported in section two of the results (i.e., after combining intrinsic motivational scales) provides support for these assertions.

Extrinsic rewards, on the other hand, are positively associated with knowledge sharing behavior. This supports the propositions asserted by social exchange theory in that workers will share their knowledge as long as it is beneficial for them to do so. This behavior will occur as a result of extrinsic rewards provided by the organization. As was found with the intrinsic motivational factors above, after combining extrinsic motivational scales, the results support these assertions.

In another attempt to confirm the relationships found and minimize the effect of relatively low scale reliabilities, a final look at the data combined knowledge creation and knowledge sharing behaviors into one scale – knowledge management behavior. These results show a positive relationship between both intrinsic and extrinsic motivational factors and knowledge management behavior. This suggests that knowledge management behavior is motivated by both intrinsic and extrinsic rewards – however, the distinction between knowledge creation and knowledge sharing behavior cannot be determined.

Limitations

This study has several limitations. The most significant limitation is the relatively low scale reliabilities. In an attempt to address this shortcoming, the hypotheses were tested using three different approaches that combined scales -- resulting in acceptable reliabilities. Although the data support the theories as hypothesized, additional research will be needed to develop new scales that provide higher levels of reliability.

There are other limitations as well. The data is cross-sectional and collected via the internet. All respondents were managers and professional staff and were employed at non-governmental nonprofit organizations. Although these types of staff members are generally characterized as knowledge-intensive workers, the reward structures of non-governmental organizations often differ greatly from their for-profit counterparts. This might impact their perspectives on rewards and knowledge creation and sharing activities. Additionally, one important note is the distinction between the quality of knowledge being shared versus the quantity of knowledge being shared (Jenkins, 1998). If workers are paid to contribute to a knowledge base, there is no guarantee that their contribution will be of high quality and potentially useful to other knowledge workers in the future. So although this study supports the proposition that extrinsic motivational factors will increase knowledge sharing, the quality of the knowledge shared was not examined. Therefore, these results should be interpreted with caution.

Implications

The results of this study suggest that organizations need to carefully design challenging work to maximize the intrinsic rewards associated with both knowledge creation and knowledge sharing. As Maslow and Herzberg suggest, employees are motivated by challenging work that helps them satisfy higher-order needs. As RBV theory suggests, this type of strategic behavior by knowledge workers can help an organization create a competitive advantage. This competitive advantage can be maintained by providing organizational opportunities for intrinsic rewards and motivation. These intrinsic reward opportunities are relatively inexpensive and can be developed through an examination and adjustment of work content such as Herzberg's motivators and the ideas of job enrichment. Through providing an opportunity for employees to fulfill higher-order needs (i.e., Maslow's top factors) through the accomplishment of their daily work, organizations can begin to establish an organizational culture that is successful at motivating strategic behavior. The results of this study suggest that through the use of intrinsic rewards, organizations can motivate knowledge creation.

The use of extrinsic rewards to motivate knowledge sharing needs to be used cautiously – with special care taken not to disrupt the intrinsic motivation involved in knowledge creation activities as suggested by Deci et al. (1999) and Jenkins (1998). When appropriately developed, extrinsic rewards help motivate employees to share their knowledge. Social Exchange theory predicts that employees will share their knowledge as long as they derive a benefit from it and this sharing will continue as long as both partners in the exchange perceive they are benefiting from it. Extrinsic rewards are one way that organizations can ensure that this mutually beneficial exchange recurs over time. The results of this study suggest that through the use of extrinsic rewards, organizations can motivate knowledge sharing. As RBV theory suggests, this type of strategic behavior by knowledge workers can help an organization sustain its competitive advantage.

Future Research

Since knowledge can be viewed as a competitive advantage, future research needs to develop an understanding of how best to motivate this strategic behavior among employees. With the complex relationship between extrinsic rewards and knowledge creation and sharing, research should investigate what type of reward systems can motivate knowledge sharing without negatively impacting creativity. Furthermore, research needs to examine possible reward systems and their effects on the quality of knowledge shared. Most importantly, future research needs to develop scales that can reliably measure the various components of knowledge management behavior such as knowledge creation and knowledge sharing.

References

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly, 25*(1): 107-136.

Barney, J. (2001). Resource-based theories of competitive advantage: A ten-year retrospective on the resource-based view. *Journal of Management*, 27: 643-650

Barney, J., Wright, M., and Ketchan, D. (2001). The resource-based view of the firm: Ten years after 1991. *Journal of Management*, 27: 625-641

Bollinger, A. & Smith, R. (2001). Managing organizational knowledge as a strategic asset. *Journal of Knowledge Management*, 5(1): 8-24.

Conner, K. & Prahalad, C. (1996). A Resource-based Theory of the Firm: Knowledge Versus Opportunism. *Organization Science*, 7(5): 477-501.

Cropanzano, R. & Greenberg, J. (1997). Progress in organizational justice: Tunneling through the maze. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology*, New York: Wiley, pp: 317-372.

Davenport, T. H., & Prusak, L. (1998). *Working Knowledge: How Organizations Manage What They Know*. Boston: Harvard Business School Press.

Deci, E., Koestner, R., & Ryan, R. (1999). A Meta-Analytic Review of Experiments Examing the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin*, 125(6): 627-668.

Gardner, D. Van Dyne, L., & Pierce, J. (2004). The effects of pay level on organization-based self-esteem and performance: A field study. *Journal of Occupational and Organizational Psychology*, 77: 307-322.

Hendriks, P. (1999). Why Share Knowledge? The Influence of ICT on the Motivation for Knowledge Sharing. *Knowledge and Process Management*, 6(2): 91-100.

Jenkins, G., Mitra, A., Gupta, N., and Shaw, J. (1998). Are Financial Incentives Related to Performance? A Meta-Analytic Review of Empirical Research. *Journal of Applied Psychology*, 83(5): 777-787.

Meso, P. & Smith, R. (2000). A resource-based view of organizational knowledge management systems. *Journal of Knowledge Management*, 4(3): 224-236.

Molm, L., Takahashi, N., & Peterson, G. (2003). In the eye of the beholder: Procedural justice in social exchange. *American Sociological Review*, 68: 128-152.

Nunnally, J. (1967). *Psychometric Theory*, *first edition*. New York: McGraw-Hill.

Peterson, R. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, 21(2), 381-391.

Selen, W. (2000). Knowledge management in resource-based competitive environments: a roadmap for building learning organizations. *Journal of Knowledge Management*, 4(4): 346-359.

Stott, K. & Walker, A**.** (1995). *Teams, Teamwork & Teambuilding: The manager's complete guide to teams in organizations*. New York: Prentice Hall.

Tampoe, M. (1996). Motivating knowledge workers – the challenge for the 1990s, in P. S Myers (Ed). *Knowledge Management and Organizational Design*, Boston, MA: Butterworth-Heinemann, pp: 179-190.

Wilkinson, H., Orth, C., & Benfari, R. (1986). Motivation Theories: An Integrated Operational Model. *Sam Advanced Management Journal*, Autumn: 24-31.

Wright, P., Dunford, B. & Snell, S. (2001). Human resources and the resource based view of the firm. *Journal of Management*, 27: 701-721.

Tables

Table 1

Survey Items and Scale Reliabilities

Data Analysis – Section One

I. Intrinsic Rewards - Knowledge Creation (KCI) (5 point scale; α =.36)

1. I find personal satisfaction in developing new ideas and insights in my area of expertise.
2. Developing new ideas and insights in my area of expertise will enhance my reputation.

II. Extrinsic Rewards – Knowledge Creation (KCE) (5 point scale; α =.77□)

1. Developing new ideas and insights in my area of expertise increases the likelihood of receiving an increase in pay or salary bonus.

2. Developing new ideas and insights in my area of expertise increases the likelihood of keeping my job.
3. Developing new ideas and insights in my area of expertise increases the likelihood of advancing my career.

III. Intrinsic Rewards – Knowledge Sharing (KSI) (5 point scale; α = .53)

1. Sharing ideas and insights with members of my workgroup or profession will enhance my reputation.
2. I find personal satisfaction in sharing ideas and insights with members of my workgroup or profession.

IV. Extrinsic Rewards – Knowledge Sharing (KSE) (5 point scale; α = .76)

1. Sharing ideas and insights with members of my workgroup or profession increases the likelihood of receiving an increase in pay or a salary bonus
2. Sharing ideas and insights with members of my workgroup or profession increases the likelihood of keeping my job
3. Sharing ideas and insights with members of my workgroup or profession increases the likelihood of advancing in my career.

V. Knowledge Creation Behavior (KC) (7 point scale; α = .61)

1. I document my ideas at work by writing them down for my own reference
2. I use my co-workers' written and/or documented ideas as a basis for forming or improving my own ideas or knowledge.
3. I use insights I've gained by working closely with others to enhance my own ideas or knowledge.
4. I merge or incorporate information documented by my workgroup or department into my own knowledge base.

VI. Knowledge Sharing Behavior (KS) (7 point scale; α = .66)

1. I formally document what I have learned for work or job-process improvement so that others in my workgroup or department may learn from my experience.
2. I research organizational or industry records to enhance my workgroup's productivity.
3. My workgroup shares knowledge with other groups inside and/or outside our organization.

Data Analysis – Section Two
Combined scales:

Intrinsic Motivation (IM) scale combines all intrinsic motivation questions from I and III above ($\alpha = .75$)
Extrinsic Motivation (EM) scale combines all extrinsic motivation questions from II and IV above ($\alpha = .89$).

IM = KCI + KSI
EM = KCE + KSE

Data Analysis – Section Three
Combined scales:

Knowledge Management (KM) scale combines all intrinsic motivation questions from V and VI above ($\alpha = .73$)

KM = KC + KS

Table 2

Correlations and Descriptive Statistics ($N = 152$)

Variables	Mean	SD	1	2	3	4	5	6	7
1. KCE[1]	5.03	1.46	--						
2. KCI[1]	6.31	.76	.369**	--					
3. KSE[1]	4.64	1.45	.884**	.379**	--				
4. KSI[1]	6.21	.82	.351**	.781**	.347**	--			
5. KC	3.76	.573	.114	.266**	.182*	.215**	--		
6. KS	3.60	.721	.283**	.276**	.331**	.188*	.503**	--	
7. IM[1]	6.26	.75	.381**	.940**	.384**	.947**	.254**	.244**	--
8. EM[1]	4.84	1.42	.971**	.386**	.970**	.360**	.152	.316**	.394**
9. KM	3.69	.55	.226**	.312**	.292*	.234**	.877**	.856**	.287**

[1] Based on a 5-point scale. All other variables are based on a 7-point scale.

$^{*}p < .05$
$^{**}p < .01$

Table 3

Summary of Regression Analysis for Variables Predicting Knowledge Creation (N=137)

Variable	*B*	*SE B*	β
KCI	.397	.136	.257**
KCE	.014	.046	.027

Note. $R^2 = .072$, adjusted $R^2 = .058$, $F(2, 135) = 5.20^{*}$.

$^{*}p < .05$
$^{**}p < .01$

Table 4

Summary of Regression Analysis for Variables Predicting Knowledge Sharing (N=137)

Variable	*B*	*SE B*	β
KSE	.148	.042	.302**
KSI	.120	.116	.089

Note. $R^2 = .117$, adjusted $R^2 = .104$, $F(2, 135) = 8.93^{**}$.

$^{**}p<.01$
$^{*}p<.05$

Table 5

Summary of Regression Analysis for Variables Predicting Knowledge Creation (N=137)

Variable	*B*	*SE B*	**β**
IM	.180	.070	.229*
EM	.018	.024	.069

Note. R^2 = .069, adjusted R^2 = .055, $F(2, 135) = 4.97^{**}$.

$^{**}p<.01$
$^{*}p<.05$

Table 6

Summary of Regression Analysis for Variables Predicting Knowledge Sharing (N=137)

Variable	*B*	*SE B*	**β**
IM	.111	.065	.149
EM	.066	.022	.262**

Note. R^2 = .119, adjusted R^2 = .106, $F(2, 135) = 9.12^{**}$.

$^{**}p<.01$
$^{*}p<.05$

Table 7

Summary of Regression Analysis for Variables Predicting Knowledge Management Behavior (N=137)

Variable	*B*	*SE B*	**β**
IM	.166	.066	.219*
EM	.072	.033	.186*

Note. R^2 = .113, adjusted R^2 = .100, $F(2, 135) = 8.57^{**}$.

$^{**}p<.01$
$^{*}p<.05$

Correspondence for this article can be sent to:

Tracy A. Hurley, Ph.D.
Texas A&M University-Kingsville, San Antonio
1400 W. Villaret Blvd.
San Antonio, TX 78224
210.413.7579
thurley@tamuk.edu

The Noncontribution of Some Data in Least Squares Regression Predictions

Teresa L. Bittner, Walden University
Kurt W. Norlin, Claremont Graduate University

Abstract

In least-squares estimation, a phenomenon can occur which has escaped attention thus far. Under certain conditions, one or more points in a data set have no influence on predictions made using ordinary least-squares models. This effect, which amounts to a discarding of *y*-data values, leads to predictions that may be suboptimal compared with predictions that use all *y*-data values.

In this paper, the relationship between data points and least squares predictions independent of those data points is demonstrated for straight-line models. When straight-line prediction is used, potentially noncontributory data can now be identified before the dependent variable data is even collected. This result can help statisticians gather their input data more efficiently and analyze existing data with better understanding.

Future research will deal with the same issue for polynomial models as well as general univariate models that are linear in the unknown coefficients.

Keywords

Least Squares, Prediction, Linear Models, Sensitivity Analysis, Noncontributory data

Introduction

The method of least squares is probably the most popular technique today to fit data to functions, estimate parameters, and determine the statistical properties of those estimates. However, a phenomenon that occurs only under certain conditions has escaped attention until now due to the fact that theoretical work generally concentrates on models and their properties rather than on individual predictions. The problem is that under certain conditions, some data fails to contribute to predictions made using least squares models. Sensitivity analysis in regression would have been the logical place for this discovery, but it is generally looked at from the perspective of changes in the parameter estimates rather than individual predictions (Chatterjee & Hadi, 1988; Cook, 1977). Even when the effect of the *i*th observation on a predicted value $(\hat{y})$ is mentioned, it is generally used to identify collinearities among regression variates, or to find data that is overly influential to the model. (Belsley, Kuh, & Welsch, 2004; Chatterjee & Hadi, 1988, p. 120–121).

Under certain conditions, the *i*th observation has no influence at all on particular $\hat{y}_j$ predictions. This finding is important because it is usually assumed that all data is being used in such predictions, and results can be skewed if data points are in fact dropping out of the prediction equations, especially for small values of *n*. It seems clear that the loss of a data point when predicting *y*-values in a linear model of the form $y = \beta_0 + \beta_1 x + \varepsilon$ is a loss of information, and such a prediction may be suboptimal in comparison to some other prediction technique that uses all the *y*-data points in its calculation. This phenomenon is also important in cases where data collection is expensive. In these cases, cost efficiency can be improved by using "dummy" data to replace real data points that would drop out of the prediction equation anyway. Further, this finding extends beyond straight line models to other models that are linear in the unknown coefficients. In fact, the number of predicted values affected increases as the number of unknown parameters increase, both in univariate and multivariate regression.

This paper describes the specific conditions under which observations drop out of prediction calculations in straight line models, develops relationships between the data point that drops out and the predicted *y*-value for which this happens, and proves the existence and accuracy of these relationships. A physical application of this phenomenon is also discussed, as are suggestions for further work on this problem.

The Phenomenon of Data Dropping Out in Least Squares Predictions

Linear modeling using the L2 (least squares) norm is a mature field of statistics. The mathematics associated with it is elegant, and the technique lends itself to many applications using closed form solutions that are efficient and convenient. However, some data has no influence on predictions made using least squares. This happens in a predictable way, and occurs for a range of models that are linear in the unknown coefficients, and in multivariate linear modeling as well. This paper will cover only the simplest linear case, while the more complex cases will be covered in future papers.

The analysis of this phenomenon begins with a simple example, followed by a short general analysis of a linear model of the form $y = \beta_0 + \beta_1 x + \varepsilon$. The implications of this analysis are then explored at length in two phases. First, we look at the special case where the *x*-values are evenly spaced in order to derive the integer cases of the phenomenon. This is followed by a full analysis of the general case. Finally, a physical application to this phenomenon is presented, and ideas for future work are suggested.

Example 1

The enrollment in Kindergarten at Allen Elementary School in San Jose, California for the last four years is as follows: (2001, 62), (2002, 43), (2003, 78), (2004, 82) (CA Department of Education, 2005). The data is plotted in Figure 1 below along with the least squares regression line.

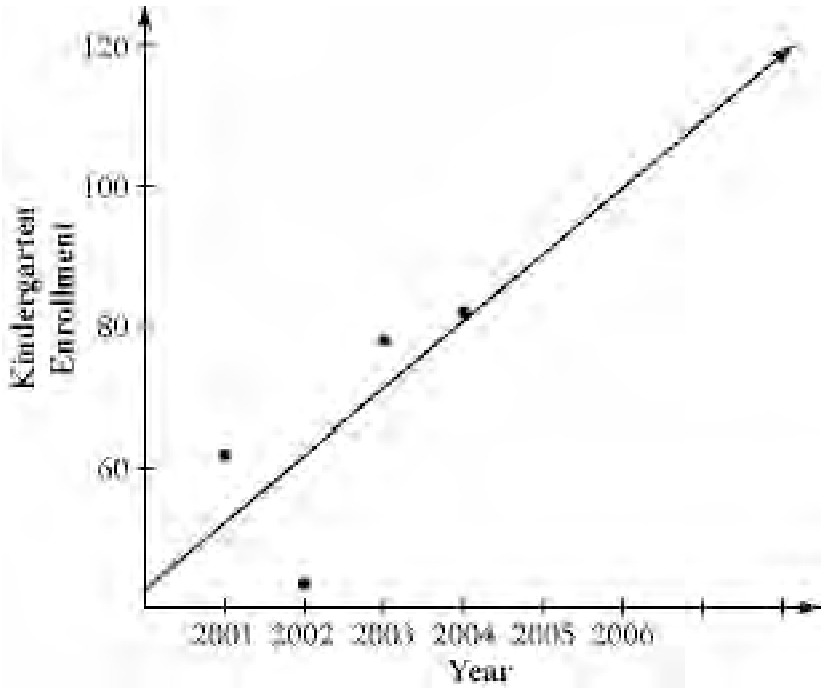


Figure 1. Allen Elementary School Kindergarten Enrollment from 2001 to 2004

Suppose the school wishes to estimate the enrollment for 2005. Assume that the enrollment behaves according to a linear function $Y = X\beta + \varepsilon$, with $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$, where the first row corresponds to 2001, the second row to 2002, etc., $Y = \begin{bmatrix} 62 \\ 43 \\ 78 \\ 82 \end{bmatrix}$, and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$. When the least squares solution is computed, the solution is $\hat{y}_i = X\hat{\beta} = X\left(X'X\right)^{-1}X'Y = 9.5x_i + 42.5$. To estimate enrollment for 2005, $x_5 = 5$ is substituted into the model to obtain $\hat{y}_5 = 42.5 + 9.5(5) = 90$. Now, suppose the second *y*-value is changed so that $y_2 = 20$ instead of the original value of 43. This changes the regression line so that $\hat{y}_i = 31 + 11.8x_i$, but $\hat{y}_5 = 31 + 11.8(5) = 90$ as before. Similarly, the second *y*-value can be changed again so that $y_2 = 120$. Now the regression yields $\hat{y}_i = 81 + 1.8x_i$. Since the value of y_2 is so large, it seems reasonable to expect that the new estimate for 2005 enrollment would be much higher than before. Yet the computation for $\hat{y}_5$ is $\hat{y}_5 = 81 + 1.8(5) = 90$ just as before. This is the case even though the regression line itself has certainly shifted. In short, it appears that the value of y_2 has no effect at all on the estimate for the 2005 Kindergarten enrollment.

It is helpful to look at this phenomenon graphically. The three regression lines, $\hat{y}_i = 42.5 + 9.5x_i$, $\hat{y}_i = 31 + 11.8x_i$, and $\hat{y}_i = 81 + 1.8x_i$ are graphed on the same set of axes along with the original data.

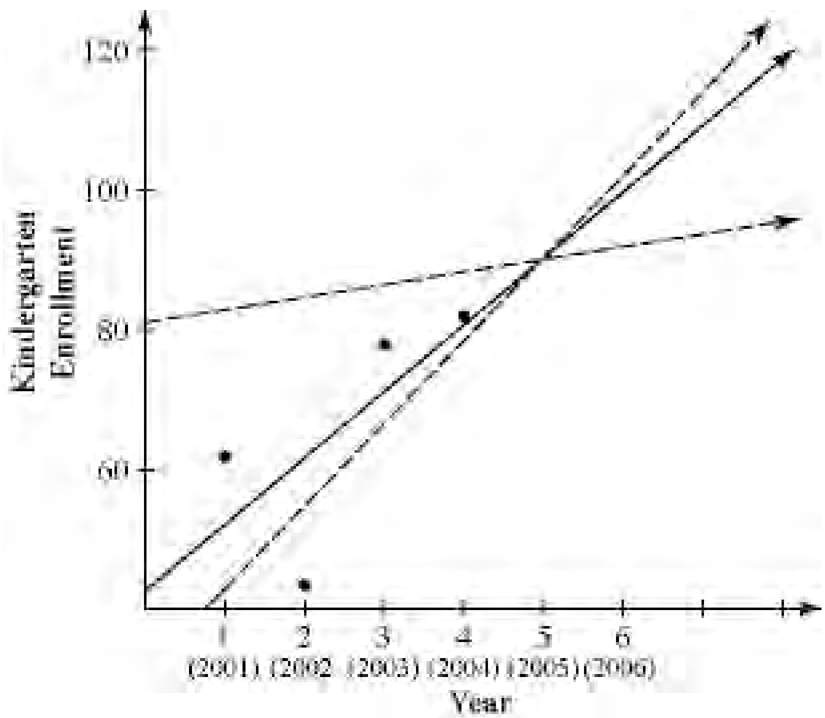


Figure 2. Allen Elementary School Kindergarten Enrollment Prediction Lines

It is easy to see in Figure 2 that the three lines intersect at (5, 90), which corresponds to the prediction that there will be 90 students in the 2005 Kindergarten class. In fact, it will later be shown that y_2 can be changed arbitrarily, and all of the regression lines will intersect at (5, 90). Notice, however, that the lines only intersect at this one point, and that for all other values of *x* the predictions for enrollment will be different when y_2 is changed. This phenomenon occurs for other values of *n* and $\hat{y}$ as well, and it will be useful to derive a mathematical relationship between noncontributory data points and the predicted values for which these points are noncontributory.

The Theoretical Basis for the Phenomenon

To set up the basis for the analytical exploration of this phenomenon, assume that the *x*-data and *y*-data values are unrestricted real numbers. In other words, assume $x_1,\ x_2,\ \cdots\ x_n$ and $y_1,\ y_2,\ \cdots\ y_n$ where the *x*-values and *y*-values are unrestricted real values.

Assume there is a linear model in *x* such that

$$\underset{n\times 1}{Y} = \underset{n\times 2}{X}\ \underset{2\times 1}{\beta} + \underset{n\times 1}{\varepsilon},$$

or, in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

It is well known that the ordinary least squares solution for this model is

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

The well known matrix form for $\hat{\beta}$ for this model is:

$$\hat{\beta}_1 = \frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \quad (1)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \quad (3)$$

and

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= \bar{y} + \hat{\beta}_1(x_i - \bar{x}).\end{aligned}$$

Now consider the special case where the x_i's are equally spaced. The simplest case is to let $x_i = i$, for $i = 1,\ 2, \cdots$ However, this simple case can be generalized to any equally spaced *x*-values by taking a linear transformation on $x_i = i$ so that $x_i = ai + b$, where *a* and *b* are scalar constants.

Note that if $x_i - \bar{x}$ is multiplied by the scalar constant *b*, then the new predicted value for *y,* named $\hat{y}_{i\,\text{new}}$, can be expressed as

$$\begin{aligned}\hat{y}_{i\text{new}} &= \bar{y} + \frac{\sum (y_i - \bar{y})(b)(x_i - \bar{x})}{\sum b^2 (x_i - \bar{x})^2} \bullet b(x_i - \bar{x}) \\ &= \bar{y} + \frac{b^2\left(\sum (y_i - \bar{y})(x_i - \bar{x})\right)(x_i - \bar{x})}{b^2(x_i - \bar{x})^2} \\ &= \bar{y} + \hat{\beta}_1(x_i - \bar{x}) \\ &= \hat{y}_i\end{aligned}$$

In other words, the scalar multiple on $x_i - \bar{x}$ does not affect $\hat{y}_i$.

Now, instead of $x_i = i$, use the more general case where $x_i = ai + b$ and examine the change, if any, in $\hat{y}_i$. Then since the only part of the equation for $\hat{y}_i$ that is affected by the transformation is the quantity $(x_i - \bar{x})$, it is sufficient to look at the effect of the transformation on this quantity. It is easily shown that

$$\begin{aligned}x_i - \bar{x} &= (a + bx_i) - \frac{1}{n}\sum (a + bx_i) \\ &= b\left(x_i - \frac{1}{n}\sum x_i\right) \\ &= b(x_i - \bar{x})\end{aligned}$$

Since the transformation only results in a scalar transformation of $x_i - \bar{x}$, and it was previously shown that a scalar multiple of $x_i - \bar{x}$ does not affect $\hat{y}_i$, this result shows that $x_i = i$ can be used without loss of generality to represent any evenly spaced *x*-values so long as the only concern is $\hat{y}_i$. The following analysis makes the assumption that $x_i = i$, but the results are valid for any equally spaced *x*-values.

The Case when n = 4

Now suppose that $n = 4$. Then $x_1 = 1,\ x_2 = 2,\ x_3 = 3,\ x_4 = 4$ is the simplest case for the x-values if they are evenly spaced. Using (1), the calculations yield

$$\sum x_i = 10,\ \sum x_i^2 = 30,\ \left(\sum x_i\right)^2 = 100, \text{ and } \hat{\beta} = \frac{1}{20}\begin{bmatrix} 30\sum y_i - 10\sum x_i y_i \\ 4\sum x_i y_i - 10\sum y_i \end{bmatrix}.$$

In order to estimate $\hat{y}_5$, the value for x_5 is substituted into the model to yield $\hat{y}_5 = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_5$, where $x_5 = 5$.

Then simplification gives

$$\hat{y}_5 = \frac{3}{2}\sum y_i - \frac{1}{2}\sum x_i y_i + 5\left[\frac{1}{5}\sum x_i y_i - \frac{1}{2}\sum y_i\right]$$
$$= -\frac{1}{2}y_1 + \frac{1}{2}y_3 + y_4$$

Interestingly, this least squares estimator for $\hat{y}_5$ is completely independent of y_2, illustrating the theory behind Example 1. Since the estimate of $\hat{y}_5$ is independent of y_2, it is clear why the various graphs of regression equations intersect at one point, even when y_2 is varied.

The Development of the General Case

Now the result for the simple linear model shown above is generalized for *n*, and the simplest case is a prediction of y_{n+1}. If the x-values are as before, then

$$\sum x_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2},$$
$$\sum x_i^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}, \text{ and}$$
$$\left(\sum x_i\right)^2 = \frac{n^2(n+1)^2}{4}.$$

The result is now generalized for $\hat{\beta}$. Using (1) gives

$$\hat{\beta} = \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2}\begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i \\ n\sum x_i y_i - \sum x_i \sum y_i \end{bmatrix}$$
$$= \frac{1}{\frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}}\begin{bmatrix} \frac{n(n+1)(2n+1)}{6}\sum y_i - \frac{n(n+1)}{2}\sum x_i y_i \\ n\sum x_i y_i - \frac{n(n+1)}{2}\sum y_i \end{bmatrix}$$
$$= \frac{12}{n^2(n+1)(n-1)}\begin{bmatrix} \frac{n(n+1)(2n+1)}{6}\sum y_i - \frac{n(n+1)}{2}\sum x_i y_i \\ n\sum x_i y_i - \frac{n(n+1)}{2}\sum y_i \end{bmatrix}$$

Therefore,

$$\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix} = \begin{bmatrix}\frac{2(2n+1)}{n(n-1)}\sum y_i - \frac{6}{n(n-1)}\sum x_i y_i \\ \frac{12}{n(n+1)(n-1)}\sum x_i y_i - \frac{6}{n(n-1)}\sum y_i\end{bmatrix}, \quad (4)$$

and

$$\hat{y}_i = \frac{2(2n+1)}{n(n-1)}\sum y_i - \frac{6}{n(n-1)}\sum x_i y_i + x_i\left(\frac{12}{n(n+1)(n-1)}\sum x_i y_i - \frac{6}{n(n-1)}\sum y_i\right) \quad (5)$$

Now a general result is sought for arbitrary *n*. In general, (4) and (5) give

$$\hat{y}_{n+1} = \frac{2(2n+1)}{n(n-1)}\sum y_i - \frac{6}{n(n-1)}\sum x_i y_i + (n+1)\left(\frac{12}{n(n+1)(n-1)}\sum x_i y_i - \frac{6}{n(n-1)}\sum y_i\right)$$

$$= \sum\left(\frac{-2(n+2)}{n(n-1)} + \frac{6}{n(n-1)}x_i\right)y_i \quad (6)$$

It can be seen from equation (6) above that the integer cases where y_i drops out can be found when the coefficient of y_i is 0. So,

$$\frac{-2(n+2)}{n(n-1)} + \frac{6}{n(n-1)}x_i = 0$$

Therefore,

$$2n + 4 = 6x_i$$

$$n = 3x_i - 2, \ x_i = 2, \ 3, \ \cdots$$

Note that the case where $n = 1$ is eliminated because it is a trivial case.
By manipulating the above equation, a more convenient form of the sequence can be written.

When $n = 4 + 3k, \ k = 0, \ 1, \ 2, \ 3, \cdots$ then y_{k+2} drops out of the prediction for $\hat{y}_{n+1}$. In particular, substituting $n = 4 + 3k$ in the above equation for $\hat{y}_{n+1}$ and simplifying shows the result that the $(k + 2)$th *y*-data point drops out when estimating $\hat{y}_{n+1}$, as stated.

The Relationship Between $\hat{y}_p$ and y_d

To continue, it will be necessary to define some notation. The individual data points to be fitted to a linear model can be represented as ordered pairs of the form (x_i, y_i). Data points that have no influence on a prediction will be denoted (x_d, y_d), where y_d is the observed data point in the *d*th position. The prediction for which (x_d, y_d) has no influence is denoted by $(x_p, \hat{y}_p)$, where $\hat{y}_p$ is the predicted value computed by substituting *x* by x_p in the model. The indices *d* and *p* are integer valued, but x_d and x_p are real numbers, unless otherwise specified.

The simplified case where the x_i's are equally spaced helped to determine which integer values of *k* cause y_k to have no influence when estimating some $\hat{y}_i$, and to find the relationship between *k* and *i*. While the integer cases have an obvious use, it is also possible to derive a closed form relationship between x_d and x_p. In this case, *d* is restricted to be a value for which x_d exists as measured data, while x_p may be any real value. Here the x_i values are unrestricted. Supposing such a relationship exists, the derivation of the relationship between x_p and x_d follows.

Recall the solution for $\hat{\beta} = (X'X)^{-1} X'Y$, where X and Y are defined as before. Assuming that there exists a value of $\hat{y}_p$ that is not dependent on y_d, the relationship between x_p and x_d is independent of the actual values of the y-data. This can be seen by taking the derivative of $\hat{y}_p$ with respect to y_d $\left(\frac{d\hat{y}_p}{dy_d}\right)$. This literally shows the change in $\hat{y}_p$ with respect to y_d. If the observation y_d does not affect the prediction $\hat{y}_p$, then this derivative should be equal to zero. Now, the $\hat{y}_p$ values are dependent on the value of x_p and the $\hat{\beta}$-values, and these values are related by the equation $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$. However, the $\hat{\beta}$-values are linear combinations of the y_i values. Therefore, once the derivative is taken, there is no y in the expression. Even without stating the expression for $\hat{y}_p$ explicitly, we can say that taking the derivative of this value with respect to y_d will eliminate all expressions that have to do with any value of y except for y_d. At that point the only thing left is the coefficient of anything having to do with y_d, since we know based on our model that there would be no expression that is nonlinear in y_d.

Since the change in $\hat{y}_p$ with respect to y_d does not depend on any of the y-values, any values of y can be used in order to derive the relationship between $\hat{y}_p$ and y_d, and their corresponding values x_p and x_d. Therefore, in order to derive the relationship between x_p and x_d, the values of the y-vector can be varied at will without loss of generality. Visually, the desired result is the point x_p at which all the various regression lines corresponding to different values of y_d intersect when the other y-values are held constant. For this purpose any two lines will suffice, and thus y-values can be chosen for maximal convenience. Thus, let all the y-values other than y_d equal 0, and let y_d be either 0 or 1. (Thus Y is either the zero vector or the indicator function on Y at d. The indicator function is the vector that contains all zeros except for a "one" in some position d.) The other y-values do not matter for this purpose anyway, so it is easiest to let them equal 0. Thus Y is the indicator function on Y at d, denoted I_d. Then

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_d & \cdots & x_n \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ x_d \end{bmatrix},$$

and

$$\hat{\beta} = \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} 1 \\ x_d \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}.$$

Now, by performing the matrix multiplications, the individual components of $\hat{\beta}$ can be expressed in terms of x_d as

$$x_p = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - x_d \cdot n}. \quad (7)$$

By solving the equation (7) for different values of n and x_p, the same integer results as before are obtained. Namely, when $n = 4 + 3k$, $k = 0, 1, 2, \cdots$, y_{k+2} drops out when estimating $\hat{y}_{n+1}$. A theorem and the proof of this result follows.

Theorem 1. Given $y = \beta_0 + x\beta_1 + \varepsilon$, Y $(n \times 1)$, X $(n \times 2)$, specified. Let $\hat{y}_p$ (p real), be a prediction based upon $\hat{\beta} = (X'X)^{-1} X'Y$. Then there exists an integer value d, $1 \le d \le n$, and $d \ne p$, such that $\hat{y}_p$ does not depend on y_d, and the relationship between p and d is specified by

$$x_p = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - n \cdot x_d}.$$

Proof: It needs to be shown that $\hat{y}_p$ does not change when y_d is varied arbitrarily, and the other y-values are fixed.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$= \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

and

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \quad (8)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (9)$$

Also, $\hat{y}_p = \bar{y} + \hat{\beta}_1 (x_p - \bar{x})$.

Note that with some simplification,

$$x_p = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - n \cdot x_d}$$ can be written as

$$x_p = \frac{\sum (\bar{x} - x_i)^2}{n(\bar{x} - x_d)} + \bar{x} \quad (10)$$

Now, by substituting equation (10) for x_p into the equation $\hat{y}_p = \bar{y} + \hat{\beta}_1 (x_p - \bar{x})$ and substituting for (8) for $\hat{\beta}_1$ as well, the equation becomes

$$\hat{y}_p = \bar{y} + \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \left(\frac{\sum (\bar{x} - x_i)^2}{n(\bar{x} - x_d)} + \bar{x} - \bar{x} \right).$$

Since $\sum (x_i - \bar{x})^2 = \sum (\bar{x} - x_i)^2$, the equation above simplifies to

$$\hat{y}_p = \bar{y} + \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n(\bar{x} - x_d)}.$$

Expanding gives

$$\hat{y}_p = \frac{1}{n} \sum y_i + \frac{\sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \cdot \bar{y})}{n(\bar{x} - x_d)},$$

and multiplying both sides of the equation by $n(\bar{x} - x_d)$ gives

$$n(\bar{x}-x_d)\hat{y}_p = \frac{1}{n}(n)(\bar{x}-x_d)y_d + \sum(x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\cdot\bar{y})$$

If y_d has no influence on equation for $\hat{y}_p$, only the parts of $\hat{y}_p$ that involve y_d need to be calculated. It needs to be proven that these terms equal zero. Therefore, noting that $\bar{x} = \frac{1}{n}\sum x_i$ and $\bar{y} = \frac{1}{n}\sum y_i$, and eliminating all the terms not depending on y_d by writing $\sum y_i = y_d + \sum_{i \neq d} y_i$ yields

$$\begin{aligned} n(\bar{x}-x_d)\hat{y}_p &= \bar{x}\,y_d - x_d y_d + x_d y_d - \bar{x}\,y_d - \bar{y}\sum x_i + \sum \bar{x}\cdot\bar{y} \\ &= -n\bar{x}\cdot\bar{y} + n\bar{x}\cdot\bar{y} \\ &= 0. \text{ QED} \end{aligned}$$

Example 2

The finance manager of a major fast food chain suspects that the gradually increasing number of tacos sold can be usefully modeled by a linear function. She has decided to compile data on the number of tacos sold for several years in order to estimate the number of tacos likely to be sold over the next several years if the pattern continues. She knows she can compile data for the last 11 years, except that the year 3 data was irretrievably lost due to a computer crash several years ago. Therefore, she can compile data for years 1, 2, 4, 5, 6, 7, 8, 9, 10, and 11, where year 11 corresponds to last year. The finance manager has noted that it will take a considerable amount of effort to retrieve the data for the number of tacos sold each year, due to the way the data was originally recorded. Therefore, she wants to make sure that all the data she collects will be used in her estimates for years 12, 13, 14, and 15. Will any of the predicted values the manager wants be independent of any of the data values?

Using Theorem 1, it is clear that the relationship between a data point $(x_d,\ y_d)$ and any predicted value $(x_p,\ y_p)$ that is independent of that data point is

$$x_p = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - n \cdot x_d}.$$

In this case, the above equation needs to be solved four separate times for x_d, once for each of years 12, 13, 14, and 15.

The required calculations are $\sum x_i = 63,\ \sum x_i^2 = 497$, and $n = 10$. The solutions to the equation for each value of x_p are given in the table below. These values of x_p correspond to the collected data value of y_p.

Table 2. Values For Which y_d Does Not Affect $\hat{y}_p$

x_p	x_d That Does Not Affect y_p
12	4.5
13	4.8
14	5
15	5.1

Therefore, the fourth data value (when $x = 5$) will not affect the predicted value of $\hat{y}_{14}$. This might be a reason for the finance manager not to bother collating the data for that year. Admittedly, given that the fourth data value is still apparently relevant to the predictions for years 12, 13, and 15, there might still seem to be a reason to go ahead and compile the fifth data value. However, we see a hint emerging that the fifth data value may not have much effect on the predictions for years 12, 13, and 15 after all. In fact, initial findings indicate that there are "degrees of relevance" for various data points that would strengthen the case for omitting the fifth data point. The analysis of this issue is beyond the scope of this paper and will be the topic of a future paper.

A Physical Application

Interestingly, there is a physical parallel to the statistical result expressed by Theorem 1. Recall the linear model in x where

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

Consider the following physical application:

Suppose there are N x-values where the x-values are allowed to be arbitrary, (i.e., not necessarily equally spaced). For each x-value, place a point mass at the corresponding point on the number line, where all the masses are 1 unit in magnitude. Now suppose all these masses are joined by massless rod connectors to form a single rigid body that is floating in space, as seen in Figure 3 below.

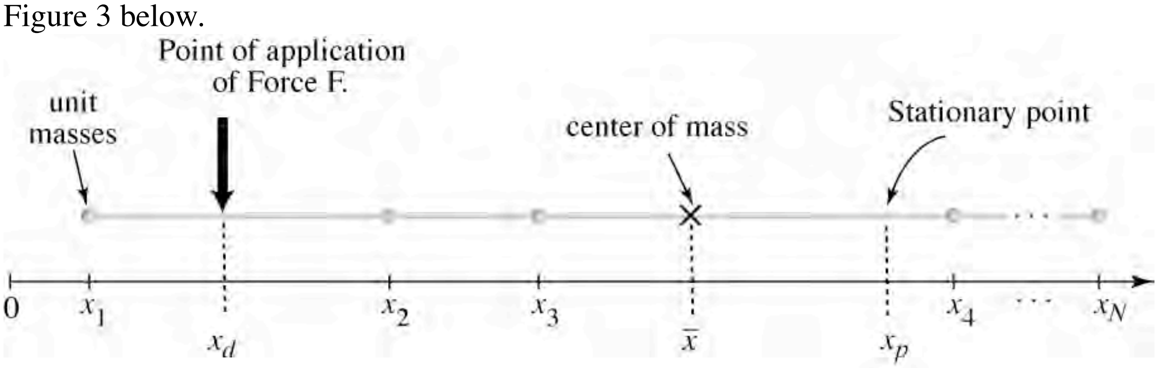


Figure 3. Point masses joined by a massless rod to form a single rigid body.

If one were to press sideways against this linear body at some arbitrary point x_d that is not the center of mass, then the body will now begin to translate and to rotate. There will, however, be a point x_p at which the effects of translation and rotation cancel out, a point that will remain stationary. The linear body will pivot about that point. (See Figure 4 below). The relationship between x_d and x_p can be derived from simple physical relationships.

We know from basic physics that

$F = ma$, where F is force, m is mass, and a is acceleration

and

$\tau = I\alpha = Fd$, where τ is torque and $I\alpha$ is the moment of inertia multiplied by the angular acceleration, all with reference to rotation about the center of mass.

For our rod-connected unit masses,

$ma = Na_{\bar{x}}$ ($a_{\bar{x}}$ = acceleration of the center of mass, N is the number of point-masses),

$Fd = F \cdot (\bar{x} - x_d)$,

and the moment of inertia is $I = \sum_{i=1}^{N} (\bar{x} - x_i)^2$.

Then $Fd = I\alpha$ becomes

$$F(\bar{x} - x_d) = \sum_{i=1}^{N} (\bar{x} - x_i)^2 \times \alpha .$$

The displacement, d_{x_q} of an arbitrary point x_q along the rod, is the displacement of the center of mass $d_{\bar{x}}$ plus the displacement due to the rod rotating an angle θ about the center of mass d_r (see Figure 4), so

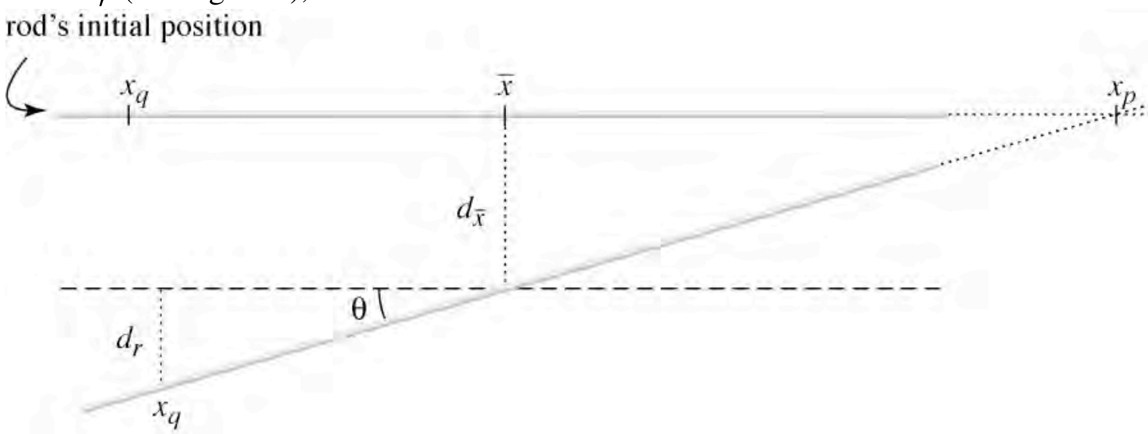


Figure 4. Pivoting massless rod.

$$d_{x_q} = d_r + d_{\bar{x}} \text{ and}$$

$$d_r = (\bar{x} - x_q)\sin\theta$$

¨Using $\theta \approx \sin\theta$ for small values of θ and differentiating twice yields

$$a_{x_q} = (\bar{x} - x_q)\alpha + a_{\bar{x}}.$$

Now, the point where $a_{x_q} = 0$ is the point x_p, which is the point that does not move when force is applied at x_d. At that point,

$$\alpha(x_p - \bar{x}) = a_{\bar{x}}$$

$$x_p = \frac{a_{\bar{x}}}{\alpha} + \bar{x}$$

$$= \frac{F/N}{F(\bar{x} - x_d) \Big/ \sum (\bar{x} - x_i)^2} + \bar{x} \quad \text{(by substitution of initial equations)}$$

$$= \frac{\sum (\bar{x} - x_i)^2}{N(\bar{x} - x_d)} + \bar{x}$$

which, after some simplification becomes

$$x_p = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - x_d \cdot N},$$

where $\bar{x}$ is the statistical notation for the physical quantity of the center of mass, x_d is the x-value that is independent of the prediction $\hat{y}_p$, and x_p is the x-value corresponding to $\hat{y}_p$.

Future Work

Preliminary results show that the result regarding data points that have no influence on predictions holds for models of the form $Y = X\beta + \varepsilon$, which are linear in the unknown coefficients but polynomial equations in *X*. Exactly *k* values of $\hat{y}_i$ will be independent of some data point for any linear model that is polynomial in *X* with power *k*. This will be the subject of a future paper. Preliminary work has also shown that the result is extendable to the general univariate model that is linear in the unknown coefficients, and for multivariate linear models as well. Future work will extend the results for these cases and compare the estimates of $\hat{y}_p$ for the least squares norm in this special case when data drops out to other methods for estimating $\hat{y}_p$.

Additionally, future work will explore the "degrees of relevance" as introduced at the end of Example 2. Finally, if we measure distance from $\bar{x}$ in one direction as positive and in the other direction as negative, then the relationship between x_p (an *x*-data point at which we want to predict y_p) and x_d (the *x*-data point at which y_d is irrelevant to that prediction) has the general shape of the function $x_p = -\frac{1}{x_d}$. That is, a data point slightly to the left of $\bar{x}$ will be irrelevant to a prediction far to the right of $\bar{x}$, and a data point far to the left of $\bar{x}$ will be irrelevant for a prediction slightly to the right of $\bar{x}$. These topics will also be covered in a future paper.

References

Belsley, D., Kuh, E., & Welsch, R. (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* New York: John Wiley & Sons.

California Department of Education, Educational Demographics Unit. District and School Enrollment by Grade, San Jose Unified School District, 2001-2004 [Electronic data retrieval]. Retrieved on September 9, 2005, from http://www.cde.ca.gov/ds/sd/cb.

Chatterjee, S., & Hadi, A. (1988). *Sensitivity Analysis in Linear Regression.* New York: John Wiley and Sons.

Cook, R.D. (1977). Detection of influential observation in linear regression, *Technometrics, 42(1),* 65–68.

The Alan Guttmacher Institute. (2004, February 19). U.S. Teenage Pregnancy Statistics With Comparative Statistics For Women Aged 20-24: Notes on Teenage Pregnancy Statistics [Electronic version], Retrieved on September 1, 2005, from http://www.agi-usa.org/pubs/teen_stats.html.

Content Management System (CMS) Implementation in a Marketing Organization of a Software Company – A Case Study

Marta Camargo, Walden University, USA

Abstract

This article is a descriptive case study of a Content Management System (CMS) initiative as implemented by the marketing department of a software company (Company A) in Silicon Valley, California. The first part explains the fundamentals of CMS and provides a theoretical framework for understanding of this technology. The second part describes the experience of Company A in both strategic and operational terms. The findings suggest that the company delivered an effective CMS solution from a technological standpoint. From an operational and managerial standpoint, however, communication stood out as a change management problem that could have jeopardized the success of the company's CMS initiative.

Keywords

Content Management System, CMS, content, publishing, authoring, content reuse, content leveraging, corporate silo, workflow, template, element, metadata, publishing network system, XML, marketing deliverables, change management.

Introduction

This case study describes and analyses the experience of Company A with CMS to determine whether the results were successful. Data analysis used a simplified version of Trochim's Theory of Pattern Matching (Trochim, 2004), to compare, contrast, and score Company A's actual practices in relation to CMS's recommended practices.

The data collection methodology for this study consisted of five types:

1. Internal process and procedure documents posted on Company A's corporate Intranet;
2. Open-ended and informal interviews with employees of Company A's marketing department;

3. Direct observation and participation (the first author worked for Company A at the time this case was initially written); and
4. Archival records on the first roll-outs and presentations from the researcher's personal archives.
5. Books and articles about CMS.

Limitations

As CMS is a relatively new technology, most sources available for research either refer to Web publishing only or are sponsored by CMS providers. As a result, this case study has an abundance of in-text citations that reference a book entitled *Managing enterprise content: a unified content strategy*, by Rockely, Kostur, and Manning (2003). This is one of the very few sponsor-free sources of information on CMS, written from a technical, operational, financial, and managerial perspective.

For technical definitions, the researchers used non-scholar sources such as www.wikipedia.com and other generic sources on the Internet. The information and definitions from these sources were validated by CMS professionals in the researcher's organization as being correct.

What is CMS?

Content Management System or CMS can be defined in many different ways. As its name suggests, it is basically a system or infrastructure that creates and manages content. A more detailed definition would be as follows:

> *A CMS is a tool that enables a variety of (centralized) technical and (de-centralized) non-technical staff to create, edit, manage and finally publish (in a number of formats) a variety of content (such as text, graphics, video, and documents), whilst being constrained by a centralized set of rules, processes and workflows that ensure coherent, validated electronic content. (Williams, 2004, ¶ 7)*

CMS has many components with definitions that are specific to the context of content management. The definitions of the most common terms used in this case study are listed below (Rockley, Kostur, & Manning, 2003).

1. "Authors" are technical writers or content experts in an organization. In a CMS infrastructure, they write content without any styling or formatting. For that, they use authoring tools that allow them to write, edit and preview their work in preparation of the final deliverables.
2. "Content" is the information companies provide about their products that is separate from its presentation. In a nutshell, content is the intellectual capital of an organization.
3. "Element" is a unique module of information that works as an information container.
4. "Metadata" is descriptive data attached to an object (element) that allows content to be indexed, retrieved, and processed.
5. "Template" carries the presentation information, including style and positioning.

6. Workflow defines the work order of the content as each element or module is produced.
7. XML is a markup language for documents containing structure information. It is short for eXtensible Markup Language. It allows authors to create their own customized tags to mark a section of a document with a formatting command.
8. XSL is a language used to create style sheets for XML. This is necessary because XML separates content from presentation. XML tags do not show where the content needs to be displayed. This is done by the XSL style sheets.

What are the benefits of CMS?

Starting with the Internet boom in the early 1990s, the Web gradually became the first place consumers went to research and purchase products and services. In 1993, there were approximately one hundred sites with a few thousand Web pages. By 2003, it was estimated that the Internet had close to two billion Web pages (Addey, Ellis, Suh, & Thiemecke, 2002), a number that continues to grow. This scenario made it imperative for companies to find better and faster ways to publish and update their Web content.

As companies looked at their content strategies to find solutions to time and cost constraints, it became apparent that although the web pages shared the same core information, there were different ways to write about it. Authors used different tools and approaches as they worked alone in their separate departments. Each writer came up with his or her own concept or presentation of the information, and each concept was isolated from the others (Rockley et al., 2003).

Companies soon realized that to be efficient and cost effective in promoting themselves, the first thing they had to do was eliminate content silos. Silos happen when organizational structures are based on the belief that content should be tailored to different target audiences and display media. The end result tends to be lack of information consistency and standardization, high costs, and absence of content leveraging mechanisms.

CMS solves this problem of inconsistency and lack of standardization by offering a unified strategy for content authoring (Rockley et al., 2003). CMS also offers solutions at an enterprise level covering content needs of different types of deliverables such as Web pages, brochures, sales guides and many others.

CMS and ROI

CMS requires efficient planning and goal-setting. Efficient planning must involve a detailed consideration of investment costs to cover the following (Rockley et al., 2003):

1. Technology acquisition of new software and hardware;
2. Training and consulting as CMS requires content and production teams to be more technologically savvy; and
3. Initial lost productivity as part of the transition process.

In view of these initial investments, the biggest challenge when it comes to making a case for adopting a CMS strategy is to develop metrics for Return on Investment reports, or ROIs (Rockley et al., 2003) to substantiate potential savings the new system can bring to organizations. The problem is that, in a way, it is almost like comparing apples and oranges, as previous costs without content reuse or leveraging and done in a silo mode belong to a different thought process and time. In this context, it not uncommon for organizations not to know or have any metrics that cover specific components such as brochures, user guides, or a single Web page, not only due to the impossibility of being that granular in their cost monitoring, but also because they were often times authored by different teams or departments that had different budgets and cost control methods.

Most CMS vendors try to help organizations calculate their ROIs by providing them with detailed and sophisticated tables of projected annual savings (Rockley et al, 2003). However, these CMS tables and cost projections are usually not accurate and can be misleading. According to Upshall (2004), these types of calculations "simply suggest a commodity-like view of CMS, whereas in practice a CMS is likely to have an effect right across the publishing process, making the figures less precise than they appear." (p. 2)

As a workaround to these problems, Rockley et al. (2003) suggest that ROI should be calculated and adjusted on an ongoing basis as the CMS system is implemented and executed. Additionally, a CMS infrastructure should only be considered as benefitial if it can definitely support an organization's business goals and objectives. For this reason, specific CMS goals need to be qualified to make sure they are consistent with the corporate vision, as well as realistically quantified as to their feasibility in terms of current costs and potential savings.

Change Management

As explained earlier in this essay, CMS is more than a system – it reflects a change in the way authors work and deliver content. Therefore, it is not uncommon to encounter resistance from authors or writers who do not quite understand why this change is necessary. These authors also have difficulties understanding the technology itself as they are usually not involved in the system development. As a result, the adoption of CMS might be embraced by some and ignored by others who just do not understand it (Robertson, 2004).

To avoid this problem, organizations should plan on their CMS implementation taking longer than anticipated to account for the time needed to overcome initial resistances to operational and cultural changes. Efforts involving internal communication activities should start in the project requirement phase and continue throughout the implementation of CMS. These activities could include, for example, the creation of a project web site or Intranet to publish status reports, strategic documents, and other information. Other resources such as an assistance hotline, face-to-face training, and informal ongoing support could also help managers get their employees' buy-in (Robertson, 2004).

The Next Level of CMS and Company A

As companies came to embrace the benefits of CMS for their Web and printed material, they also realized the possibilities of sharing content among other types of media such as video and audio. This realization happened in conjunction with new advancements in publishing software, which became more user-friendly and less costly. The end result was the birth of a new concept called the Network Publishing.

The Network Publishing concept is revolutionary in that it releases content from the constraint of being developed only for a particular medium. With it, users can now view, store, forward and print content anywhere they want (A. T. Kerney, White Paper, 2001). Figure 1 below illustrates this network publishing concept.

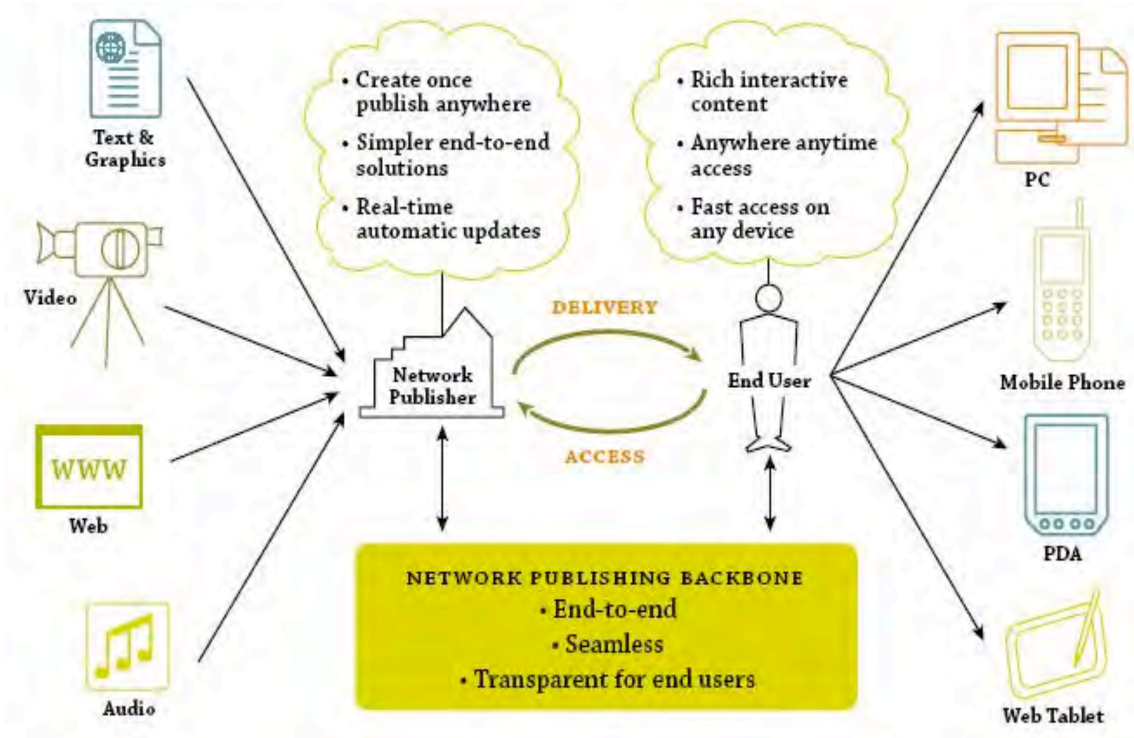


Figure 1. Network Publishing concept. From *Network Publishing–creating value through digital content*. White Paper by A. T. Kerney (2001), p. 15. Printed with permission.

Based on this new content development and publishing paradigm, Company A found itself in a privileged position as its product lines include software applications for creation, publication, and delivery of text and graphics for different types of media such as print, web, and multimedia. As a result, Company A rolled out its new vision based on a network publishing philosophy that would serve as a direction to their product offerings and strategic partnerships starting in 2002.

Company A's Marketing

In view of the nature and type of their deliverables and in support to the company's new vision, Company A's Marketing Communications (MarCom) department expressed interest in being the first department to adopt a CMS solution in the form of a network publishing infrastructure to produce the company's main marketing deliverables. The MarCom deliverables used to be developed by four separate teams:

1. The Packaging and Collateral team created boxes, brochures, datasheets, CD art, and channel copies.
2. The Merchandizing Tools team created product information spreads for sales catalogs, postcards, web banners, and print ads.
3. The Web team created the product pages published at the corporate web site.
4. The Multimedia team created presentations, videos, and demos.

Each team had different management and production teams, as well as separate budgets and vendors. Therefore, issues with consistency and costs commonly arose.

This scenario was typical of the content silos as described earlier in this case study. Without knowing what the other was doing, each team in the MarCom department was consistently creating new content and deliverables instead of leveraging from already existing ones or working with each other to create joint ideas.

To pursue a CMS strategy, MarCom's management assigned a task force to go through all their deliverables and develop a list of components that shared common content. With this information, the department created an infrastructure that defined content that could be leveraged in different deliverables.

Due to confidentiality and proprietary issues, the actual infrastructure cannot be displayed in this case study. However, Company A's content infrastructure was similar to the generic illustration shown in Figure 2 below:

Element	Metadata
Product Name	All
Product Description	
Product Desc. Short	All
Product Desc. Med	Brochure
Product Desc. Long	Brochure
Graphic	Show catalog Brochure Web site
Features	All[a]
Feature title	Brochure Press release
Feature item	All
Feature item	
Feature item	
Feature item	
Feature item	
Benefit item	Brochure Press release Web site[a]
Tag line	Brochure Press release Web site

Figure 2. CMS infrastructure example. From *Managing enterprise content: a unified content strategy*. White Paper by The Rockley Group Inc. (2003), p. 12. Printed with permission.

Each element on the left column of Figure 2 identifies common content units that appear in different marketing deliverables as shown under the metadata column. Figure 3 and Figure 4 that follows show examples of how these elements become actual content in those deliverables.

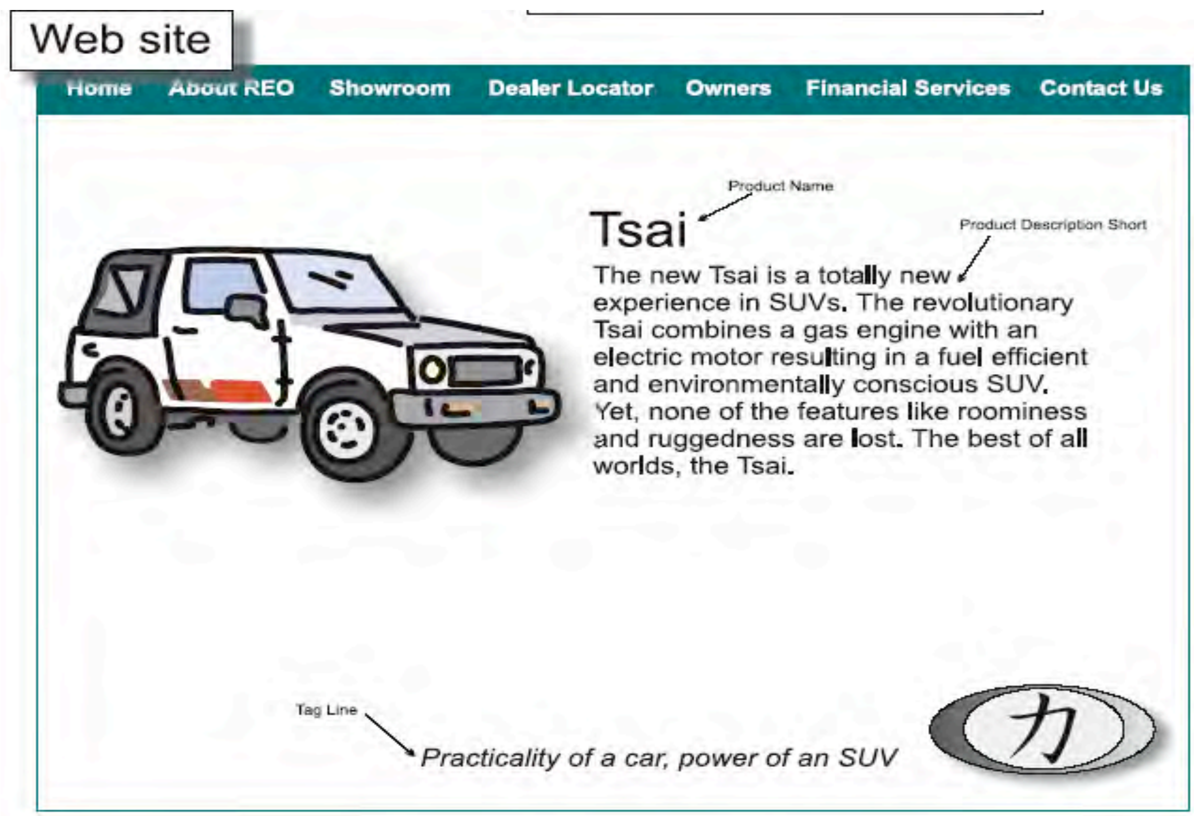


Figure 3. Metadata output example. *Managing enterprise content: a unified content strategy*. White Paper by The Rockley Group Inc. (2003), p. 17. Printed with permission.

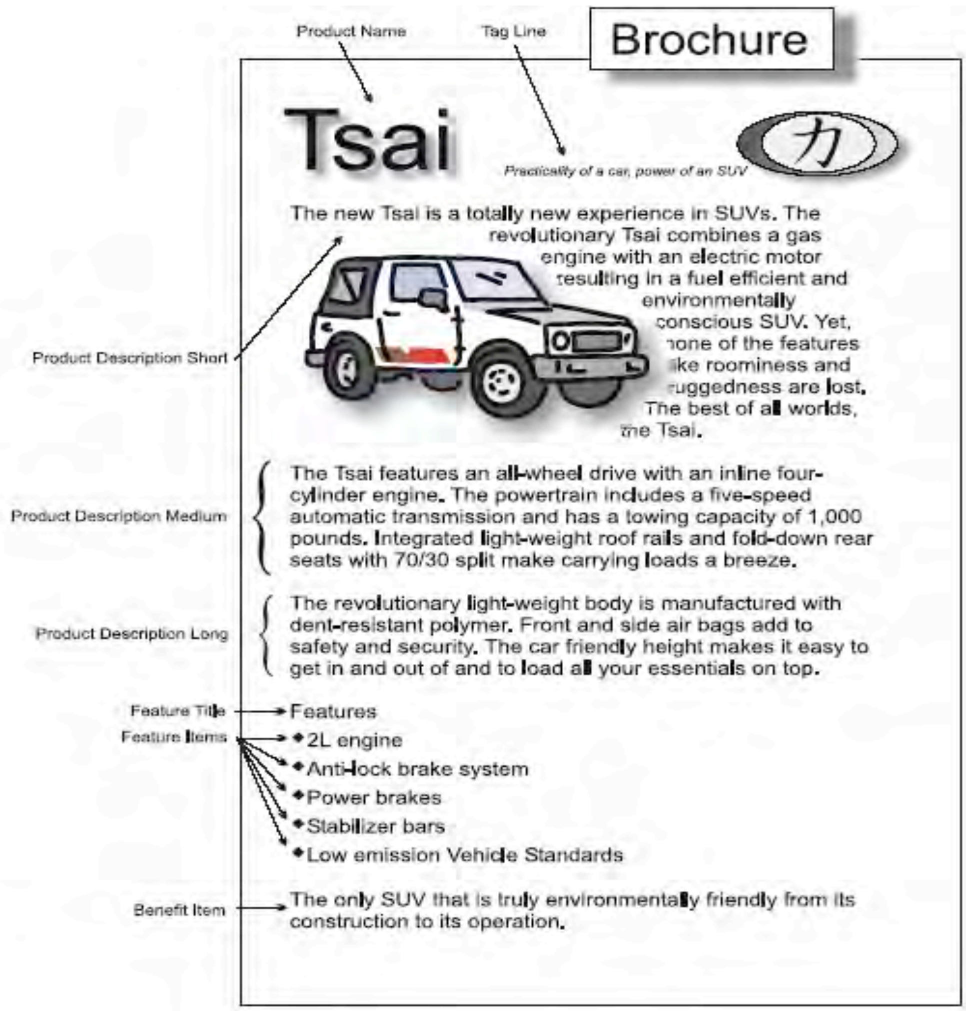


Figure 4. Metadata output example–Brochure. *Managing enterprise content: a unified content strategy*. White Paper by The Rockley Group Inc. (2003), p. 17. Printed with permission.

The CMS infrastructure was supported by authoring processes that relied on Company A's own software products as authoring tools to create text and graphics for their marketing deliverables. Company A also developed a review workflow system based on the PDF technology that allowed reviewers to read the content, make changes, and approve it still in a metafile format. This sped up considerably the turnaround time for final approvals as all deliverables were processed by the system simultaneously, as opposed to being created and approved one by one in the final stages of formatting.

Company A also used their own products in the final layout and delivery in different types of media such as print and web. For that, Company A's system used structured text files that contained all the product information as an eXtensible Markup Language (XML) file. Using a proprietary and internally developed server application, this XML file was bound to an eXtensible Stylesheet Language Transformation (XSLT) template to produce outputs in the form of templates that became marketing deliverables such as the product box, datasheet, and Web pages.

ROI

Although Company A's Executive Team provided full support to the introduction of a network publishing initiative, MarCom's management team still had to submit ROI reports to justify new hardware and software acquisition, as well as staff realignment (i.e., firing and hiring). The problem was that as a silo mentality was predominant among the various groups that produced content, it was difficult to ensure much accuracy in estimations and predictions regarding savings with CMS.

Unfortunately, MarCom's management team presented a ROI report with unrealistic expectations pointing to immediate savings to the Executive Team, which is what Rockley et al. (2003) warned as a dangerous pitfall. Consequently, one year later, after the first network publishing projects were completed, actual costs revealed a break-even scenario (at the most), instead of significant short-term savings as predicted.

In any case, as this new system supported the strategic goals and vision of the company, the ROI fiasco became more of a management issue, not a technological one. In terms of technology and efficiency, the system proved to work, producing an effective CMS infrastructure.

Company A's Change Management Practices

After the network publishing was confirmed as the standard system for marketing content creation and production, MarCom's management had to realign internal resources to meet the new requirements of an automated system. From developing marketing content and deliverables internally and individually on a per launch basis, the department changed to an automated content management system that mass-produced content. The new system had specific technical requirements that called for a new unified infrastructure. As a result, teams were laid-off and a new way of doing things was introduced.

Interestingly, with so many changes and details involved in the transition to a new way of doings things, management found itself so caught up in the technical challenge of the system that little time was set aside to communicate their vision and goals properly. Most employees felt that they did not understand the big picture and were not sure how their roles could fit into this new structure.

Although some documentation was available to explain the new process and system, MarCom's management did not adequately explain or define the new roles and responsibilities of team members. Teams had to figure out by themselves what they needed to do to perform their tasks. The network publishing system and subsequent organizational changes were quickly rolled-out to middle managers, who did not understand the technical complexities well enough to explain the system to their own people. As a result, to most people in the MarCom organization who were interviewed for this case study, lack of communication made the network publishing solution come across as much more complicated than it actually was.

Similarly to what Robertson (2004) described, and as pointed out earlier in this case study, while some teams understood what they needed to do, others simply could not follow it. Additionally, MarCom's management did not create an Intranet site or any other communication channel to clarify issues during the transition to the new system. Project and functional managers had to figure out processes and procedures by themselves in order to train internal and external personnel resources.

The first project was a pilot in December 2002 that met with mixed results. The creation and review milestones went well, but the output phase showed some problems with the quality of the templates and automation. Throughout 2003, more projects went through the system cycle; with each one, engineers moved in to fix bugs and improve each of the components of the system. Eventually, by 2004, all main launches became network publishing system projects with all vendors and internal resources fully proficient in all aspects of the system.

In spite of the communication flaws attributed to MarCom's management during the roll-out and transition phases, most employees now believe that this CMS solution is a success not only in delivering content faster, but also in making all production and creative teams work cohesively in the same workflow.

Final Analysis

To summarize and wrap up all points discussed in this case, Figure 5 shows a simplified and high level adaptation of Trochim's (2004) Theory of Pattern Matching. Each Theoretical Realm item based on the CMS theory has a fixed score of 10. Utilizing this theory, each Observational Realm based on the discussed CMS practice was scored on a scale from 1 to 10, depending on its corresponding match with the theory (i.e., 10 would be the highest match and 1 would be the lowest). Five randomly selected MarCom employees were invited to validate this analysis and final score.

Theoretical Realm		Score		Observational Realm
Unified strategy (elimination of content silos)	►	10	◄	Unified strategy (elimination of content silos)
Consistency	►	10	◄	Consistency
Development Time	►	10	◄	Development Time
Technology	►	10	◄	Technology
ROI	►	5	◄	ROI
Communication	►	3	◄	Communication

Figure 5: Scoring with Trochim's Theory of Pattern Matching

This quick scoring based on Trochim's Theory of Pattern Matching technique shows Company A's system as an overall effective technological solution. On the other hand, the low scores for Communication indicated a need for better change management practices. Lack of communication led to rumors and resistances to change due to confusion and rework. All this affected productivity and almost jeopardized the success of the new system.

Conclusion and Recommendations

As a descriptive case study, the focus of the narrative was on describing the situations and events that surrounded the implementation of a CMS infrastructure in a marketing organization of a software company–Company A. With that goal in mind, data showed that as a technology, Company A succeeded in delivering a CMS-based publishing solution for their content. However, from a change management perspective, data showed that Company A failed to recognize the value of effective communication as a key element in their organizational change.

In that regard, Rockley et al. (2004) explained that any company engaging in similar CMS initiatives should always develop a communication plan to document the ongoing status of the change including specific information related to the successes achieved in each phase of the implementation, as well as problems that were encountered and their solutions. Additionally, Rockley et al. provided the following recommendations that Company A and other companies should consider before engaging in CMS initiatives:

1. Develop a transition plan based on a phased approach with usability test and specific metrics to evaluate results.
2. Define and document new roles such as Enterprise Project Coordinators, Information Technologists, and Information Architects as early as possible in the change process.
3. Involve change agents or experts to help communicate and implement the change.

References

Addey D., Elllis J., Suh P., & Theimecke D. (2002). *Content management systems: tools of the trade*. Birmingham, UK: Glasshaus.

Managing enterprise content: a unified content strategy. White Paper (2003). (n.d.). The Rockley Group Inc. Retrieved November 10, 2004, from http://www.rockley.com/whitepap.htm.

Network Publishing–creating value through digital content, White Paper (2001). (n.d.). A.T. Kerney. Retrieved October 10, 2004, from http://www.Company Acom/aboutCompany A/pressroom/pressmaterials/networkpublishing/pdfs/netpubwh.pdf

Robertson, J. (2004). Successfully deploying a Content Management System. *KM Column*. Retrieved November 10, 2004, from http://www.steptwo.com.au/papers/kmc_deployingcms/.

Rockley A., Kostur, P., & Manning S. (2003). *Managing enterprise content: a unified content strategy*. Indianapolis, IN: New Riders.

Trochim, W. M. (2004). *Research methods knowledge base*. Retrieved, October 10, 2004, from http://www.socialresearchmethods.net/kb/.

Upshall, M. (2004). CMS—is it worth it? *Bookseller*, 5142, 22-24. Retrieved November 5, 2004, from Business Source Premiere.

Williams, S. (2004). What is a Content Management System or CMS? *Content Management Site*. Retrieved November 25, 2004 from http://www.contentmanager.eu.com/history.htm.

Technology Management in Services: Understanding of Sources of Value Creation and Service Output Characteristics

Hyungu Kang, Central Michigan University, USA

ABSTRACT

Technology has had a profound effect on how services are designed and delivered. We have seen many changes emerging in services as a result of technology. A survey of the existing classification schemes of services shows that they are not designed specifically for the management of technology in services, and, thus, falls short when applied to such an environment. This paper presents the argument that a classification scheme, which incorporates the idea of the source of value added and the degree of product component delivered, offers an effective method of classifying services for the management of technology. This paper also explores the strategic technology management issues on different types of service.

Keywords

Customer value, Service characteristics, Service Classification, Service technology, Technology management

INTRODUCTION

In recent years, much has been written about the growth and importance of the service sector. As this area of the economy has grown, the role of technology as a source of building competitive advantage has been critical. For nearly a decade, students of the service sector have been exploring the relationship between technology and competitive advantage of service sector (Kandampully, 2002; Sirilli and Evangelista, 1998; Kellogg and Nie, 1995; Alic, 1994; Morone and Berg, 1993; Hayes and Thies, 1991; Quinn, Baruch and Paquette, 1998; Vitale, 1986; Berg, 1973). We also have seen many changes emerging in services as a result of technology. Numerous examples of technological advances in banking (Morone and Berg, 1993; Vitale, 1986), health care (Schooleman, 1993), and professional service (Empson, 2001; Behara, 2000; Fischer, 1996), education (Thukral, 1995; Lin, 1995; Lee, 1995) have been central to industry-wide growth and the creation of new markets. For example, Ernst & Young offers continuing education to its national workforce through a private satellite network. A cost- and time-effective alternative educational delivery system provides tremendous opportunities to educate both internal staff and the firm's clients, and almost eliminates non-billable travel time and travel costs. American Airlines created distinctive competitive advantages through differential expertise using Sabre system and SMARTS over the years. Merrill Lynch's Cash Management System, based on a Banc One information system that integrated a variety of customer accounts, gave the company an enormous and sustainable market share advantage over fast following competitors that offered similar services. FedEx solidified its competitive standing with a system for tracking the status of customers' packages in real time. Citibank dramatically increased its market share in retail banking in part through the introduction of automatic teller machines well before its competition. Over time, it has become increasingly critical for firms in these and other service industries to effectively manage the use and adoption of critical technologies.

The differences between different types of services have presented a particular difficulty for researchers attempting to offer prescriptive advice regarding effective management practices. Because of the broad range of service types, researchers in the area have worked to develop methods of classifying services in a number of ways: the nature of the service product (Sasser, Olsen and Wyckoff, 1978; Schmenner, 1986), the view of the service recipients (Alic, 1994; Hill, 1977; Chase, 1981), the standpoint of the main producer of the service products (Thomas, 1978), synthesized classification (Kotler, 1980; Lovelock, 1983), and the time and space utility creation (Hsieh and Chu, 1992). These classification schemes have been useful in providing insight into effective management practice. But developing classification schemes is not enough. If they are to have managerial value, they must offer strategic insights.

While these classification schemes provide useful guidance from many perspectives, they were not designed explicitly to be applied to technology management in services. Therefore, they often have their shortcomings. For example, estimating the degree of customer contact provides little guidance on effective technology implementation practices in banking. That is why it is important to develop ways of classifying services that highlight the characteristics they have in common, and then to examine the implications for technology management. This paper presents the argument that a conceptual framework, which incorporates the idea of the source of the value added and the degree of product component delivered offer an effective method of classifying services for the management of technology.

In the text below, we begin by providing an overview of several of the popular classification schemes of services. We then outline the important role of technologies in service. Finally, we outline the utility of existing frameworks of services and present the argument for the new perspective as an effective means of managing technologies in services.

CHARACTERISTICS AND CLASSIFICATION OF SERVICES

Early research in the field sought to differentiate services from standard manufacturing environments in a number a ways; focusing particularly on four generic differences—intangibility, heterogeneity, perishability of output, and simultaneity of production and consumption (Sasser, Olsen and Wyckoff, 1978; Quinn, 1992; Fitzsimmons and Fitzsimmons, 1994; Shostack, 1977; Rathmell, 1974; Bateson, 1977; Zeithaml, Parasuraman and Berry, 1985; Edgett and Parkinson, 1993). Although these characteristics are still commonly cited, they have been criticized as too generic to understand the processes underlying service delivery (Verma and Boyer, 2000; Cook, Goh and Chung, 1999; Lovelock, 1991; Lovelock and Yip, 1996), and there is growing recognition that they are not universally applicable to all services. Lovelock and Yip (1996) provide an alternative set of eight characteristics. These characteristics begin with the nature of the output—performance rather than an object—and also include customer involvement in production, people as part of the service experience, greater likelihood of quality control problems, difficulty in customer evaluation, lack of inventories for services, greater importance of the time factor, and availability of electronic channels of distribution. These characteristics provide a useful starting point for thinking about the distinctive aspects of service management.

From an operational perspective, Lovelock and Yip (1996) assign core services to one of three broad categories (people-processing, possession-processing, and information-based services), depending on the nature of the process and the extent to which customers need to be physically present during service production.

People-processing services involve tangible actions to customers in person. These services require that customers themselves become part of the production process, which tends to be simultaneous with consumption (i.e., passenger transportation, health care, food service, lodging services). Possession-processing services involve tangible actions to physical objects to improve their value to customers (i.e., freight transport, warehousing, equipment installation and maintenance, car repair, laundry, and waste disposal). The object needs to be involved in the production process, but the customer does not, since consumption of the output tends to follow production. Finally, information-based services involve collection, manipulation, interpretation, and transmission of data to create value (i.e., accounting, banking, consulting, education, insurance, legal services, and news). In production of information-based services, customer involvement is often minimal.

The core service product is typically accompanied by a variety of supplementary service elements (i.e., order taking, billing, and payment), which, taken together are commonly referred to as the "service bundle". These supplementary elements not only add value, but also provide the differentiation that separates successful service providers from the rest.

According to Thomas (1978), the traditional perspective of the service business that the service is "invariably and undeviatingly personal, as something performed by individuals for other individuals" is erroneous. Automatic car washes, automated banking services, and computer time-sharing are just three examples of the many service businesses in which the service is primarily delivered through interactions with automated equipment (Collier, 1985; Collier, 1983). The strategic requirements for these businesses are obviously quite different from those in which individuals perform services for other individuals.

Thomas suggests one way to separate service businesses into general types (equipment-based and people-based service), according to different strategic management requirements. According to Thomas (1978), to effectively manage a specific service business, it is necessary to answer two questions: (1) how is the service rendered? (2) what type of equipment or people renders the service?

Thomas argues that service business managers can analyze their companies and take advantage of the strategies that are uniquely available to them. Also, based on the above classification of service type, management in service businesses can better understand the nature of the strategic opportunities of their firm's position. As will be seen later, the model we present for the management of technology in services builds on these ideas of Thomas, expanding on and refining the equipment/people based idea and incorporating issues specific to technologically rich environments.

In sum, a survey of the existing classification schemes of services shows that they are not designed specifically for the management of technology in services, and thus fall short when applied to such an environment.

TECHNOLOGY MANAGEMENT IN SERVICES

Technology as a source of building competitive advantage is playing an important role in many service industries. Service technologies typically are described as “knowledge technologies” (Perrow, 1967; Thompson, 1967), and defined by Dubin (1968) as “the body of ideas which express the goals of the work, its functional importance, and the rationale of methods employed.”

Implementing technology in service environments creates many difficulties for the firm, as traditional manufacturing-based methods may not be appropriate. According to Mills and Moberg (1982), this is because applying “the manufacturing transformation process model” (in which raw materials are processed and can be put into inventory as finished goods waiting for customer demand) to service operations is problematic. Because services cannot be stored, one is forced to conceive of input and output sectors as depositories of client/customers waiting for entry on the one hand and exits on the other. The conversion is restricted to the direct interaction between client/customer and service worker. Yet, this neglects several important features of service operations. First, technologies may be employed outside the conversion (interaction) process itself. Thus, technology is applicable at all stages of the service production system, not just within the so-called conversion process. Second, the transactional nature of the interaction within the conversion processes invites recycling, skipping, and aborting. Thus, the conversion process may be composed of several typical information-exchange subprocesses, but these constitute a reciprocal rather than a linear sequence. Third, because of the interactive nature of many services, unintended or serendipitous conversions may occur at any stage in the system. Client/customers may experience insight, inspiration, or even instantaneous learning with the workflow. In sum, the distinction among inputs, conversion, and outputs is muddied in the service production process and it often bears little resemblance to a strict production environment.

The split between manufacturing and service-based industries often has far reaching effects. In a discussion of the diffusion of new technologies in services associated with ‘technology push’ and ‘demand pull’ processes, Barras (1986) outlines the familiar ‘product cycle’ theory of innovation, whereby successive waves of innovation in a particular technology shift progressively from product innovations which generate new devices to process innovations which improve the quality or performance of existing devices. However, based on an empirical study of the adoption of information technology in service industries, Barras suggests that an opposite process of innovation tends to operate, which can be termed a 'reverse product cycle'. Thus even the very nature of the evolution of product and process development may be different.

The context within which knowledge technologies are brought to bear on client problems also is different (Quinn, 1992). To produce a service, the customer and service worker must interact for the delivery of the service to be complete (Mills and Margulies, 1980). Thus, the client/customer and service worker exchanges information and commitment (‘Transaction Process’ in Thompson). The necessary contact between client/customer and service worker has several other implications. First, the spectrum of customer reactivity is a striking outcome of many interactions. Second, service technologies require a high capacity for information processing within the technical core. Finally, service technologies depend on the client/customer not only to provide the information that constitutes the raw material to be worked on, but also often to make use of client efforts in the service production process (Thukral, 1995). This means that client/customer is involved directly in the production of his/her own wants.

Thus, in summary, service operations seem to have several features that distinguish them from manufacturing. These emerge from the intangibility of service outputs and the peculiar nature of service technologies and the service process, including perishability. Unfortunately, there is not much to be learned regarding technology implementation in services. Thus firms are left with no choice but to follow guidelines established in manufacturing settings found in the existing literature.

As was noted earlier, because of the wide breadth and diversity of services in existence, researchers have tried many ways of "splitting up" the service sector (Verma and Boyer, 2000; Cook, Goh and Chung, 1999; Chase, 1981; Thomas, 1978; Kotler, 1980; Lovelock, 1983; Shostack, 1977; Rathmell, 1974; Lovelock, 1991). Services vary significantly across a number of dimensions, and thus must be managed differently in different scenarios. One would not expect to find "universal" approaches to the effective management of every type of service. Technology—viewed as encompassing skills, expertise, know-how, and the organization of work—has become a major competitive weapon in service sectors (Harvey, Lefebvre and Lefebvre, 1997; Lewis, 2002; Alic, 1994; Lovelock, 1983; Fitzsimmons and Fitzsimmons, 1994; Clemons, 1986; Feeny and Ives, 1990; Guile and Quinn, 1988). To gain a competitive advantage, technology is often implemented in ways that yield improvements in the bundled end product. As with other management practices, the process of implementing a new technology in services differs significantly from industry to industry within the sector. In some types of services, the product, the production process, or both, have a significant dependence on the implementation of an advanced technology. Digital computers, for example, have become integral to the production of new types of services (i.e., data processing itself and computer-assisted architectural drafting). In these environments, however, computers are used to enhance people's skills rather than replace them entirely. The operator (a draftsman) still needs basic skills and knowledge of drafting in order to use the computer-assisted equipment effectively. Such a technology might render some skills obsolete (drawing, lettering…), but at its core, the technology is designed to *enhance* an operator existing knowledge base.

In another service environment, advanced technologies may have a prominent place in the product/process environment in a way such that it reduces the required skill set of the operator, instead embedding some of the skills and knowledge into the equipment itself. Consider the adoption of automated cooking equipment at a fast food restaurant. Rather than enhance the operators skills, the technology adoption here would effect the core of the "value added" in the service bundle. Thus, it is the technology that provides the value-added, rather than the worker. For example, since 1990, Arby's has been installing automated systems for food selection and payment as part of its plan to revolutionize customer service. The roast beef sandwich chain is using Touch 2000, a computerized system that allows customers to order food from a touch-sensitive display screen, processes the information through an IBM PS/2 Model 30 computer, and displays it on IBM monitors in the food preparation area. The chef, while working within a service operation, uses the technology (i.e., an automated french-fry cooker) to deliver a bundled service to the customer. In such environments, the competencies of the technology itself go a long way in creating the competitive advantage.

As can be seen from these examples, technology implementation in services in which a considerable amount of the value-added comes from the worker him/herself is different from those where the value-added is in the equipment. The implementation process in such environments differs greatly. For example, worker buy-in and feedback would be more critical where the technology is enhancing worker skills. Even the rationale for implementing the service in the first place would be quite different. Fischer (1996) studied new audit technology adoption in CPA firms, where the technology serves to enhance the knowledge and skills of the accountant. He found that the benefits of new technology in this particular service sector did not result directly from the adoption and use of the new technology. Rather, these benefits resulted from the concomitant reduction or elimination of other audit procedures that had been performed in the past.

The core tenets of Advanced Manufacturing Technology (AMT) adoption literature (Davis, 1996) may not be appropriate in these environments because the portion of the service which is knowledge based is inherently tacit, difficult to measure and difficult to transfer. The quality of the service is not "conditioned into the consumer", and the professional will stand or fall with the perception of the customer. The result is that the quality of the service is not as routinized. New technology acceptance in these cases is at the discretion and description of the individual distributing it, and the abilities of the individual delivering the service will ultimately drive its final quality. In addition, while service outcomes in these areas can often be measured through traditional quality measures, the customers' perception of the service outcome may vary considerably over time.

Additionally, services can vary significantly in the amount of physical product delivered in the bundle. Those with a significant product component behave differently than those where a "pure" service is delivered. In "pure" services, the degree of customer contact (Chase, 1978) is often higher, and perceptions play a greater role. In services with a significant product component ("goods-like"), traditional methods of controlling quality might come into play. For example, some restaurants (i.e., Diners) have poor or even rude service, but thrive because of the high quality/value of the food delivered. While this distinction has been made in previous research (and thus will not be elaborated on here), it remains central to effective technology management in the service sector.

A SERVICE CLASSIFICATION

Sources of Customer Value

As the above discussion outlines, in technology management in services, it is useful to divide the service sector according to the utilization of those services that have the majority of knowledge embedded in the service production system (i.e., fast food, automated car wash, broadcasting, theaters, and museum) versus those services which are based on the knowledge of the point person providing the service (i.e., teaching, exercise clinics, computer graphic, professional consulting, and legal services). We classify *Knowledge-based Services* as those services in which the majority of customer value is provided by the knowledge of the person providing the service. We classify *Knowledge-embedded Services* as those services that embed the customer value in the system that provides the service. As was noted earlier, existing classifications of service business (i.e., Thomas, 1978; Lovelock, 1983; Hsieh and Chu, 1992; Lovelock and Yip, 1996) do not show the process and/or mechanism of customer value creation. In view of the important relationship between customer value and service, this perspective focuses on how customer value is created, what mechanisms are utilized in "value-added", and how competence in service can be enhanced. An important difference is that this framework focuses on where the "value added" comes from, rather than what the process "looks like".

Characteristics of Service Output

Services can be classified to the extent to which physical product is incorporated within the output. Those services which have a substantial product component (*'Goods-like' Services*) are quite different from those which do not (*'Pure' Services*). In the technology adoption and implementation phase, a service, which has a significant product component, may, in many ways, behave like a production environment. For example, the similarities between a production facility and the flow of product through a fast food service provider are obvious. As such, many issues (such as type and method of employee training and empowerment) learned through technology implementation in a production environment would be applicable here. Furthermore, the driving reasons behind technology adoption in this service with a significant product component might again resemble those found in a production environment: efficiency, conformance quality, flexibility, yield, etc. Contrast this with a more "pure" service, where the technology will most likely interact directly with the customer, as opposed to those occurring "behind the scenes", (i.e., call waiting, pay-per-view movies); these technologies are more often implemented to directly provide the customer with new features and capabilities, and are less concerned with production-based improvements in yield, etc.

Table 1. A Classification of Services

	Knowledge-Based Services	Knowledge-Embedded Services
Goods-like Services	Type I Computer graphic, Computer Aided Design, Beauty salons, Exercise clinics, Haircutting	Type II Automated car washes, Fast Food, Passenger/Freight transportation, Laundry, Dry cleaning, Vending machines
Pure Services	Type III *Education, Professional services, Legal services, Health care, Information services, Management consultants, Accountants*	*Type IV* Package delivery, Shipping and distribution, Broadcasting, Telephone operator, Security services, Banking/Insurance, Theaters/Museums, Travel/Recreation

Type I: Knowledge-based 'Goods-like' Services

Consider the implementation of highly sophisticated computer graphic equipment (hardware and/or software) in a firm that provides computer graphic design services for art, publication, video production, and presentation. While computer graphic equipment is a powerful tool, the technology is typically aimed at enhancing the existing skills and capabilities of the design workforce, rather than replacing it. Adopting the state of the art graphic design system does not eliminate the need for workers with knowledge and training in structural design. Rather, the technology enables the user to gain benefits such as increased speed (efficiency), the ability to create a "library" of designs to be re-used (repeatability), and increased accuracy of drawings (quality). This reflects the fact that this service has a significant product component—the drawing itself. As will be seen in later examples, the above mentioned goals are not shared with technologies implemented in pure services.

Since the primary "value-added" of the computer graphic service remains in the hands of the designer, we would classify this as a knowledge-based service. As such, the choice of technology and method of training would reflect the needs of the service to closely match the requirements of the individual worker. It might require close monitoring and, perhaps, customization of training, as well as significant worker buy-in to change. In selection of both technology and workforce, much effort might be needed because this kind of service requires high skill levels (as measured by years of education) and relatively high pays. Well-selected and implemented technology gives service workers a capability to customize the customer's various needs. This places computer graphic equipment implementation in Type I of Table 1.

Type II: Knowledge-embedded 'Goods-like' Services

Contrast this with the typical implementation of automated technology in an automated car wash. Many of the goals are the same; increased efficiency, repeatability, better conformance quality, etc. This is because the two are similar in that there is a significant product component in what is delivered: a drawing and a meal. However, the role of the person interacting with the technology is very different. In the above computer graphic example, the technology was an extension of the highly educated or experienced worker—enhancing their skills and building off of their knowledge. In the restaurant example, the situation is very different—the worker is now just an extension of the technology, and the technology has reduced or replaced the need for certain worker skills. As such, one could argue that the value-added is embedded in the technology, not the worker, and it is the technology that is enabled by the worker (by pushing the appropriate buttons when signaled, etc.). With technology, the workforce can standardize their work and service. This would be in Type II of Table 1.

Because of these differences, the implementation of technology in this knowledge-embedded environment would be managed differently too. As discussed above, while advanced technologies may have a prominent place in the product/process environment, in general, neither the nature of the service nor the nature of the production process is affected by the technology in a fundamental way. For example, worker training in this environment would most likely be quite routine and standardized, and it is unlikely that the worker (car wash operator) would have significant input regarding the specific piece of equipment purchased. Additionally, there is less need to select a workforce with high skill levels and educational requirements. In such implementations, system-wide technology adoption/implementation strategies that can be learned from the experience of manufacturing sector might be useful.

Type III: Knowledge-based 'Pure' Services

Consider the implementation of new communication technology now becoming popular in education that allows classes to be taught via satellite or Internet. Clearly, this is a knowledge-based service. The teacher/instructor/professor remains central to the process, and is providing the basis for the class. As such, worker (instructor) buy-in, training, and involvement in the process are keys to successful implementation. The new medium technology provides an extension of the instructors' capabilities, enabling him/her to reach students in locations that would not be possible through traditional methods. It is critical that the instructor knows the abilities and limitations of the technology in order to effectively use it. Often continuing learning/training might be needed. As in the case of computer graphic service, selection of appropriate technology for service might be critical.

However, unlike our previous two examples, there is not a significant product component to this service. This significantly affects the technology implementation process. First, the goals of the technology are different from traditional production based improvements such as conformance quality and efficiency. The main benefit of this technology directly addresses the needs of a set of new customers who do not live in close proximity to the classroom. Rather than operating "behind the scenes" like in our computer graphic and automated car wash examples, in this case, the customer has significant interaction with the technology itself. In fact, the customer might have chosen to purchase the service (the class) on the basis of the offerings of the technology. The technology itself is designed to create attributes that the customer is aware of and finds desirable. Thus, implementing technology in such pure service environments, the customer himself would be more at the forefront in the selection process. Further, both the instructor (worker) and the student (customer) would need training in how to use the new technology. This differs from the above examples, where the customer might be oblivious to the technology itself.

Type IV: Knowledge-embedded 'Pure' Services

Our final illustration of the utility of the framework uses the example of package delivery (i.e., UPS, FedEx). Again, it is the technology that provides the main source of competitive advantage here. The mapping and order tracking technologies are the backbone of the elaborate delivery network. As in our example in the automated car wash, the worker in this environment (the driver and sorter) relies heavily on the technology to guide their daily activities. The worker responds to the technology's direction (i.e., creating routes), rather than the other way around. As such, technology implementation issues such as training and worker buy-in would resemble those in Type II, with customization and worker feedback playing a relatively limited role. The major focus of the adoption/implementation of new technology in this kind of service might be the way of system-wide competence enhancing in production process of service.

Unlike the automated car wash, however, firms operating such as package delivery operating in Type IV do not have a significant product component in their service. As a pure service, the customer interacts with the process (and often the technologies—i.e., customer on-line tracking). As a result, the customer tends to play a greater role in the process as a whole. The customer interacts with the service provider to select the delivery options, and intangibles, such as driver courtesy, greatly affect customer satisfaction. As a knowledge-embedded service the "production system" can be standardized through new technology, but as a pure service the issue of interaction with customer might be a key factor in the technology implementation process and success.

DISCUSSION AND IMPLICATIONS

Widespread interest in the technology adoption/diffusion of the service sector among both academics and practitioners is a relatively recent phenomenon. This might reflect the fact that technology management in the service sector has significantly lagged behind that in the manufacturing sector. Service technologies have radically altered the strategic environment in ways that offer significant new opportunities and threats for service businesses (Guile and Quinn, 1988; Quinn and Paquette, 1990; Quinn, Baruch and Paquette, 1987; Skinner and Chakrabarty, 1982). In addition, service technologies offer a variety of ways for service providers to add value within their own operations. Thus, the adoption of innovative technology in the service sector can have widespread impact on the value-added activities for a service firm or industry.

This paper proposes a classification scheme which, incorporating the idea of the source of the "value-added" and the degree of product component delivered, offers an effective method of classifying service for the management of technology. In view of the important relationship between customer value and service, this perspective focuses on how customer value is created, what mechanisms are utilized in "value-added", and how competence in service can be enhanced.

Each type of this classification scheme has important managerial and technological issues. In a service which has characteristics of significant product component and provides a customer value by service workforce (Type I), the adoption/implementation of new technology is aimed at enhancing effectiveness, improving existing skills and capabilities of the workforce, and does not eliminate the need for workers with knowledge and experience. The choice of technology and method of training would reflect the needs of the service to closely match the requirements of the individual worker. It might require close monitoring and, perhaps, customization of training, as well as significant worker buy-in to change.

As in the case of implementation of computer graphic equipment, new technology adoption in Type II (i.e., automated car wash, fast food) is aimed at increasing efficiency, improving quality, and reducing costs. As an example, Boston Chicken invested heavily in an information-management system. This system began at the point-of-sale (POS) terminals, which doubled as training aids. The system provided managers, staffers, and headquarters with instant communication with other parts of the network. In addition, the in-store computer system, Intellistore, helped store managers schedule labor, track inventory, and manage production. Intellistore also featured forecasting functions that help the manager estimate purchasing, staffing, and preparation requirements. However, in this kind of knowledge-embedded service, the workforce typically has a low skill level and is an extension of the technology. The technology has replaced the need for worker skills. Technology adoption/implementation is focused here on the competency enhancing abilities of the system, not on the capability of workforce. In this environment, technology implementation issues such as training and worker buy-in would be routine.

In a service which is knowledge-based and 'Pure' service component (Type III), the customer has significant interaction with the technology itself. The customer might have chosen to purchase the service on the basis of the offerings of the technology, and the customer himself would be more at the forefront in the selection process. In this environment, the technology would again be implemented differently. The new technology provides an extension of the service provider's capabilities, perhaps enabling him/her to reach more customers or provide unique product attributes that would not be possible through existing methods, as in the case of distance education. Because the worker is central to the value-added, worker buy-in, training, and involvement in the process are key to successful implementation. At the same time, traditional manufacturing goals such as quality improvement and cost reduction are not as common.

Like package delivery in our classification scheme, some services are both knowledge-embedded and 'Pure' service component (Type IV). Like our example in the automated car wash, the worker in this environment relies on the technology to guide their daily activities. In addition, technology implementation issues such as training and worker buy-in would resemble those in the case of automated car wash. However, in this environment, training customization and worker feedback play a relatively limited role. However, unlike Type II, as a pure service, the customer interacts with many of the service technologies and might be involved in the selection of technology, or at least its critical attributes.

The framework presented in this paper is designed to be a starting point upon which to build theory and establish effective management practice in each of the four types. The classification scheme as proposed provides essential managerial insights. It provides a richer framework in which to manage technology in services in that it incorporates two aspects of services (the source of core value-added and the degree of product component) critical to technology implementation. Therefore, it allows a better understanding of the nature of service within a technological context. As such, it incorporates the importance of technological innovation in service production, the factors underlying the process of customer value creation, and the means of enhancement of service production system, providing managers in technologically rich service industries a more comprehensive tool than currently available.

REFERENCES

Alic, J.A. (1994). Technology in the service industries, *International Journal of Technology Management*, 9(1), 1-14.

Barras, R. (1986). Towards a theory of innovation in services, *Research Policy*, 15(4), 161-173.

Bateson, J. (1977). Do we need service marketing. In Eiglier, P., Langeard, E., Lovelock, C.H. and Bateson, J. (Eds.) *Marketing Services: New Insights*, Marketing Science Institute, Massachusetts.

Behara, R.S. (2000). Process innovation in knowledge-intensive services. In J.A. Fitzsimmons and H.J. Fitzsimmons (Eds.), *New Service Development*, Sage Publications, Inc., Thousand Oaks: CA.

Berg, S.V. (1973). Determinants of technological change in the service industries, *Technological Forecasting and Social Change*, 5, 407-426.

Chase, R.B. (1981). The customer contact approach to services: Theoretical bases and practical extensions, *Operations Research*, 29(4), 698-706.

Chase, R.B. (1978). Where does the customer fit in a service operation?, *Harvard Business Review*, 56(6), 137-142.

Clemons, E. (1986). Information systems for sustainable competitive advantage, *Information and Management*, 11(3), 131-136.

Collier, D.A. (1983). The service sector revolution: The automation of services, *Long Range Planning*, 16(6),10-20.

Collier, D.A. (1985). *Service Management: The Automation of Services*, Reston Publishing Company, Virginia.

Cook, D.P., Goh, C., & Chung, C.H. (1999). Service typologies: A state of the art survey, *Production and Operations Management*, 8(3), 318-338.

Davis, J.E. (1996). Citibank ambushed by ATM networks, Fortune, May 13, pp. 34.

Dubin, R. (1968). *Human Relations and Administration* (3rd edition), Prentice Hall, New Jersey.

Edgett, S. & Parkinson, S. (1993). Marketing for service industries: A review, *The Service Industries Journal*, 13(3), 19-39.

Empson, L. (2001). Introduction: Knowledge management in professional service firms, *Human Relations*, 54(7), 811-847.

Feeny, D. and Ives, B. (1990). In search of sustainability: Reaping long term advantage from investment in information technology, *Journal of Management Information Systems*, 7, 27-46.

Fischer, M.J. (1996). Real-izing' the benefits of new technologies as a source of audit evidence: An interpretive field study, *Accounting, Organizations and Society*, 21(2, 3), 219-242.

Fitzsimmons, J.A. & Fitzsimmons, M.J. (2004). *Service Management: Operations, Strategy and Information Technology*, McGraw-Hill, New York.

Fitzsimmons, J.A. & Fitzsimmons, M.J. (1994). *Service Management for Competitive Advantage*, McGraw-Hill, New York.

Guile, B.R. & Quinn, J.B. (1988). *Technology in Services: Policies for Growth, Trade, and Employment*, National Academy Press, Washington, D.C.

Harvey, J., Lefebvre, L.A. & Lefebvre, E. (1997). Flexibility and technology in services: A conceptual model, *International Journal of Operations & Production Management*, 17(1), 29-45.

Hayes, R.M. & Thies, E.A. (1991). Management of technology in service firms, *Journal of Operations Management*, 10(3), 388-397.

Hill, T.P. (1977). On goods and services, *Review of Income and Wealth*, 23, 315-318.

Hsieh, C. & Chu, T. (1992). Classification of service businesses from a utility creation, *The Service Industries Journal*, 12(4), 545-557.

Kandampully, J. (2002). Innovation as the core competency of a service organization: The role of technology, knowledge and networks, *European Journal of Innovation Management*, 5(1), pp. 18-26.

Kellogg, D.L. & Nie, W. (1995). A framework for strategic service management, *Journal of Operations Management*, 13, pp. 323-337.

Kotler, P. (1980). *Principles of Marketing*, Prentice-Hall, New Jersey.

Lee, M. (1995). Leading the way: College and campuses provide an insight into how technology may change the nature of instruction, *Wall Street Journal*, Nov. 13, pp. R28.

Lewis, M.A. (2002). Selecting and implanting service technology: Control, uncertainty and competitive advantage, *The Service Industries Journal*, 22(2), 17-42.

Lin, A. (1995). Far out: Distance education is more than a camera in a classroom, *Wall Street Journal*, Nov. 13, pp. R28+.

Lovelock, C.H. (1991). *Services Marketing*, Prentice Hall, New Jersey.

Lovelock, C.H. (1983). Classifying services to gain strategic marketing insights, *Journal of Marketing*, 47(3), 9-20.

Lovelock, C. H. & Yip, G. S. (1996). Developing global strategies for service business, *California Management Review*, 38(2), 64-86.

Mills, P.K. & Margulies, N. (1980). Toward a core typology of service organizations, *Academy of Management Review*, 5(2), 255-265.

Mills, P.K. & Moberg, D.J. (1982). Perspectives on the technology of service operations, *Academy of Management Journal*, 7(3), 467-478.

Morone, J. & Berg, D. (1993). Management of technology in the service sector: Practices in the banking industry, *Journal of High Technology Management Research*, 4(1), 123-137.

Perrow, C. (1967). A framework for the comparative analysis of organizations, *American Sociological Review*, 32, 196-208.

Quinn, J.B. (1992). *Intelligent Enterprise*, The Free Press, New York.

Quinn, J.B., Baruch, J.J. & Paquette, P.C. (1988). Exploiting the manufacturing-service interface, *Sloan Management Review*, 257(6), 45-56.

Quinn, J.B., Baruch, J.J. & Paquette, P.C. (1987). Technology in services, *Scientific American,* 257(6), 50-58.

Quinn, J.B. & Paquette, P.C. (1990). Technology in services: Creating organizational revolutions, *Sloan Management Review*, 31(2), 67-78.

Rathmell, J. (1974). *Marketing in the Service Sector*, Winthrop Publishers Inc., Massachusetts.

Sasser, W.E., Olsen, R.P. & Wyckoff, D.D. (1978). *Management of Service Operations: Text, Cases, and Readings*, Allyn & Bacon, Massachusetts.

Schmenner, R.W. (1986). How can service businesses survive and prosper?, *Sloan Management Review*, 27(3), 21-32.

Schooleman, S. (1993). Service, technology positions Pyxis for success, *Health Industry Today*, 56(6), 16-17.

Shostack, G.L. (1977), Breaking free from product marketing, *Journal of Marketing*, 41(2), 73-80.

Sirilli, G. & Evangelista, R. (1998). Technological innovation in services and manufacturing: Results from Italian surveys, *Research Policy*, 27, pp. 881-899.

Skinner, W. & Chakrabarty, K. (1982). *The Impact of New Technology: People and Organizations in Service Industries*, Pergamon Press, New York.

Thomas, D.R.E. (1978). Strategy is different in service businesses, *Harvard Business Review*, 56(4), 158-165.

Thompson, J.D. (1967). *Organizations in Action*, Basic, New York.

Thukral, I.S. (1995). *Quality in Services through Design: An Integrated Approach,* Unpublished doctorate dissertation, Rensselaer Polytechnic Institute, New York.

Verma, R. & Boyer, K.K. (2000). Service classification and management challenges, *Journal of Business Strategies*, 17(1), 5-24.

Vitale, R. (1986). The growing risks of information systems success, *MIS Quarterly*, 10(4), 327-334.

Zeithaml, V.A., Parasuraman, A. & Berry, L.L. (1985). Problems and strategies in service marketing, *Journal of Marketing*, 49(2), 33-46.

Index of Authors