CHAPTER 2
PROBLEM FORMULATION AND PROPOSED SOLUTION METHODOGY

2.1 PROBLEM FORMULATION

Since Drucker [Drucker 1954] first put forward the concept of human resource (HR), great progress has been made in the theory of human resource management (HRM). The competitiveness of an organization essentially depends on HR. For middle to large scale organization, a medium to long term planning of HR is crucial to the overall success of the organization. One issue of interest in the HR planning is relevant to the age distribution. In general, young workers or employees are good at their energetic characteristics, fresh knowledge from academic institutions, less expenditure in terms of salary, etc. Senior employees, though not physically strong as the young employees, have advantages in their long experiences and capability in solving critical situations. However, the senior employees put harder burdens on the organization expenditures. Regarding to the intellectual capital, the experiences of the knowledge workers at the senior levels are indispensable to the competitiveness of the organization (see [Stewart 1997] and [Edvinsson and Malone 1997]). To be sustainable and viable, the organization needs to have an age distribution in a balanced aspect, i.e. an appropriate age distribution.

In real situations, it is possible that the age distribution which exists in an organization requires a adjustment into a desired age distribution. The adjustment of the present age distribution to the desired age distribution can be achieved in several ways. The first problem of interest herein thus introduces a systematic adjustment in the age profile that the HR department may consider as an approach for the modification of an existing or a present age distribution. The adjustment in the age distribution is proposed to be performed in an evolutionary manner in which the number of employees is selectively adjusted from year to year within a prescribed time frame. The adjustment of the number of employees is carried out by adding

adjustment magnitudes to the existing number of employees at the selected age groups. The determination of the respective adjustment magnitudes is formulated in forms of a constrained optimization problem, where the adjustment magnitudes are the variables to be optimally determined. It is noted that there exist criteria for recruitment or lay-off at every selected age group. These criteria can be defined in terms of theoretical knowledge, practical knowledge, experiences, etc. In addition, the criteria are naturally not identical for the age groups but they must be considered fair when comparing between different age groups. The definitions of the criteria for the adjustment in each age group is, however, not within the scope of the present research and thus will not be the subject of further discussion hereinafter.

The HRM policy for an age distribution is stated as follows. For a given age distribution at present year $P_0 = \{E_{A1}(t_0),\ldots,E_{Aj}(t_0),\ldots,E_{ANAge}(t_0)\}$, it is desired that the age distribution $P_{NYear} = \{E_{A1}(t_{NYear}),\ldots, E_{Aj}(t_{NYear}),\ldots,E_{ANAge}(t_{NYear})\}$ in the next *NYear* years after the present year t_0 be close to the desired age distribution $P_D = \{E_{A1}{}^{D},\ldots, E_{Aj}{}^{D},\ldots,E_{ANAge}{}^{D}\}$ as much as possible. The adjustment of the age distributions P_0 to P_{NYear} is achieved via the consecutive adjustment in the number of employees at various ages. Accordingly, a mathematical expression which represents such a procedure can be given as

$$E_{A_j}(t_i) = E_{A_{j-1}}(t_{i-1}) + \delta_{A_j}(t_i) \quad ; \; i = 1,\ldots,NYear \text{ and } j = 1,\ldots,NAge \qquad (2.1)$$

$$E_{A_0}(t_i) = 0 \quad ; \forall i \qquad (2.2)$$

$$A_j - A_{j-1} = 1 \quad ; \forall j \qquad (2.3)$$

and $$t_i - t_{i-1} = 1 \quad ; \forall i \qquad (2.4)$$

, where $E_{Aj}(t_i) \in \Re$ and $\delta_{Aj}(t_i) \in \Re$. Note that (2.3) says the difference between the consecutive age group is equal to 1 year old. The number of employees $E_{Aj}(t_i)$ is taken as percentage of total number of employees, i.e.

$$\sum_{j=1}^{NAge} E_{Aj}(t_i) = 100 \quad ; \forall i \qquad (2.5)$$

This adjustment has to be decided and implemented by the HRM department and is the subject of the study in this paper. When $\delta_{Aj}(t_i)$ equals to 0's for all A_j's and t_i's, (1) is just an age evolution for every passing year. Consequently, (2.1) describes both the dynamics of age evolution and the process of the age-distribution adjustment. Equation (2.1) may be extended to the case of resignation. However, it is beyond the scope of this study.

According to the HRM policy for an age distribution above, the following optimization problem is formulated:

$$\underset{\delta_{Aj}(t_i)}{Min}\ ERR = \sum_{j=1}^{NAge}\left(E_{Aj}^{D} - E_{Aj}\left(t_{NYear}\right)\right)^2 \quad ; i = 1,\ldots,NYear \text{ and } j = 1,\ldots,NAge \qquad (2.6.1)$$

, or $$\underset{\delta_{Aj}(t_i)}{Min}\ ERR = \sum_{j=1}^{NAge}\left(E_{Aj}^{D} - E_{Aj}\left(\delta_{Aj}\left(t_i\right)\right)\right)^2 \quad ; i = 1,\ldots,NYear \text{ and } j = 1,\ldots,NAge \qquad (2.6.2)$$

, subject to $$E_{Aj}\left(t_i\right) \geq 0 \quad ; \forall i \text{ and } \forall j \qquad (2.7)$$

and $$\sum_{j=1}^{NAge} E_{Aj}\left(t_i\right) = 100 \quad ; \forall i \qquad (2.8)$$

where

NAge : The total number of age groups

Nyear : The year that the adjustment in the age distribution is expected to meet the desired age distribution

P_0 : The age distribution at present year

P_{NYear} : The age distribution at the *NYear*th year after the present year

P_D : The desired age distribution

A_j : The *j*th age group

t_i : The *i*th year

$E_{Aj}(t_i)$: The number of employees in the age group A_j at time t_i

E_{Aj}^{D} : The desired number of employees in the age group A_j

$\delta_{Aj}(t_i)$: The adjustment magnitude of the number of employees in the age group A_j at time t_i

$\Re$: The set of real numbers

ERR : The total discrepancy between P_{NYear} and P_D

$\delta_{Aj}(t_i)$, $i = 1,\ldots,NYear$ and $j = 1,\ldots,NAge$, are the variables to be optimally determined. A set of inequality constraints (2.7) are required for the non-negativity of the number of employees at all times. The 100-percentage criteria of the number of employees at all times, as given previously by (2.3), are preserved by a set of equality constraints (2.8). The optimization problem considered here involves the dynamics of a system subjected to multiple constraints.

For several years now, skills management or competency management has been suggested as a way to more effectively utilize employee skills in the work place. The concept originated from HRM as a way to align HR processes (like selection, appraisal and development) to job requirements and organizational strategy. Moreover, it has been suggested in Knowledge Management (KM) that approaches defining competencies can support knowledge management processes like goal setting and evaluation, or the assignment of teams in knowledge-based organization. In both approaches "skills" or "competencies" are being defined in organizations in order to describe characteristics of individual employees to make better use of their expertise or to develop it further. Competencies are "internal capabilities that people bring to their jobs. They maybe expressed in a broad, even infinite, array of on-the-job behavior". Competencies can be defined as personal characteristics (knowledge, skills, and abilities) of employees which are relatively stable across different situation and also in terms of distinguishable elements of underlying capacities or potential which allow job incumbents to act competently in certain situations. Competency development is such that people acquire new competencies predominantly in interaction with real job situations and tasks. New competencies are being developed when a person enters a new situation or task in which action is not predetermined. Reflecting on the outcome or receiving feedback from a more experienced person helps in this development. This view is in accordance with a large body of research

showing the importance of informal learning as opposed to formal training when it comes to learning at the workplace. The variety of competency development methods can be used for informal learning, including mentoring, coaching, networking, modeling, effective leadership and interactions in a team environment.

Human Resource Development (HRD) supports organizational competency in term of improvement. The idea of improvement overarches all HRD definitions, models and practices. To improve means "to raise better quality or condition: To make better". HRD is also a title which represents the latest evolutionary stage in the long tradition of training, education and developing people for the purpose of contributing towards the achievement individual, organizational and societal objectives. The second problem consider in this research thus aims at the issue of HRD.

The continuous improvement in the competencies of knowledge workers in this study is assumed to be carried out in terms of education. The graduates are input to the work system after their graduation. The knowledge workers who resign and are retired from the system are also taken into account. The description and assumptions of the problem including the corresponding mathematical model are as follows:

1. Total budget here is for educating students to be knowledge workers.
2. The time in producing the knowledge workers from students takes 4 years.
3. First batch of students starts their studies in the university at the present year, i.e. the 0^{th} year.

The long-term planning of knowledge workers is proposed in terms of the minimization of total of the differences between supply and demand. Correspondingly, the minimization is formulated as follows:

$$\text{Minimize} \quad O(\boldsymbol{S},\boldsymbol{D}) = \sum_{i-1}^{NYear} (SW_i - DW_i)^2 \qquad (i = 1,...,NYear) \qquad (2.9)$$

where,

$\boldsymbol{S}$: Vector of supplied knowledge workers in the ith year SW_i

$\boldsymbol{D}$: Vector of demanded knowledge workers in the ith year DW_i

SW_i : Supplied knowledge workers in the *i*th year

DW_i : Demanded knowledge workers in the *i*th year

NYear : Total number of years to be considered in the planning

, where $$\boldsymbol{S} = \left[SW_1 \quad \cdots \quad SW_i \quad \cdots \quad SW_{NYear}\right]^T \tag{2.10}$$

$$\boldsymbol{D} = \left[DW_1 \quad \cdots \quad DW_i \quad \cdots \quad DW_{NYear}\right]^T \tag{2.11}$$

Subject to: $$0 \le TC_i\left(\boldsymbol{S},\boldsymbol{D}\right) \le TBU_i \tag{2.12}$$

TBU_i : Upper bound of total budget available in the *i*th year

TC_i : Total cost in the *i*th year

The evolution of the students at respective academic year in each year is:

$$ST2_i = \alpha_{1i} ST1_{i\text{-}1} \tag{2.13}$$

$$ST3_i = \alpha_{2i} ST2_{i\text{-}1} \tag{2.14}$$

$$ST4_i = \alpha_{3i} ST3_{i\text{-}1} \tag{2.15}$$

$$G_i = \alpha_{4i} ST4_{i\text{-}1} \tag{2.16}$$

$ST1_i$: number of first academic-year students attending the universities in the *i*th year

$ST2_i$: number of second academic-year students in the *i*th year

$ST3_i$: number of third academic-year students in the *i*th year

$ST4_i$: number of fourth academic-year students in the *i*th year

G : number of graduates from the university in the *i*th year

α_{1i} : percentage of the first academic-year students that pass to the second year in the *i*th year

α_{2i} : percentage of the second academic-year students that pass to the third year in the *i*th year

α_{3i} : percentage of the third academic-year students that pass to the fourth year in the *i*th year

α_{4i} : percentage of the fourth academic-year students that graduate from the university in the *i*th year

The supply is described by

$$SW_i = SW_{i-1} + G_i - RW_i - RTW_i \quad (2.17)$$

In this optimization problem, the supplied knowledge workers in the *i*th year, i.e. SW_i (i = 1,...,*NYear*) are the design variables to be determined. The knowledge workers who resign and are retired from the work system in each year are given by

$$RW_i = \alpha_{RWi} SW_{i-1} \quad (2.18)$$

$$RTW_i = \alpha_{RTWi} SW_{i-1} \quad (2.19)$$

in which,

α_{RWi} : percentage of RW_i with respect to $SW_{i\text{-}1}$

α_{RTWi} : percentage of RTW_i with respect to $SW_{i\text{-}1}$

RW_i : number of knowledge workers resigning from the work system in the *i*th year

RTW_i : number of knowledge workers retiring from the work system in the *i*th year

The annual cost incurred by the education is

$$TC_i = FCOST_i * ST1_i + SCOST_i * ST2_i + TCOST_i * ST3_i + FTCOST_i * ST4_i \quad (2.20)$$

where

$FCOST_i$: cost for educating a first year student in the *i*th year

$SCOST_i$: cost for educating a second year student in the *i*th year

$TCOST_i$: cost for educating a third year student in the *i*th year

FTCOSTi : cost for educating a fourth year student in the *i*th year

Finally, the initial conditions are

$$SW_0 = SW_o \tag{2.21}$$

$$ST1_0 = ST1_o \tag{2.22}$$

$$ST2_0 = ST2_o \tag{2.23}$$

$$ST3_0 = ST3_o \tag{2.24}$$

$$ST4_0 = ST4_o \tag{2.25}$$

in which

SW_o : Supplied knowledge workers in the *starting* year

$ST1_o$: number of first academic-year students in the *starting* year

$ST2_o$: number of second academic-year students in the *starting* year

$ST3_o$: number of third academic-year students in the *starting* year

$ST4_o$: number of fourth academic-year students in the *starting* year

It should be noted that the number of the first academic-year students attending the universities in respective years is the variable to be determine, i.e. $ST1_i$ ($i = 1,\ldots,NYear$).

The third problem of interest is the management of knowledge workers under limited resources and time. Such an application can be found in the appointment systems of outpatient service. Healthcare is a fundamental and essential factor in daily life. High demand of outpatient service is persistently high. It is thus necessary to provide medical personnel as well as facility to satisfy a prescribed level of service requirement. From an economic point of view, cost of outpatient service has tendency of increasing every year whereas financial support to the service may not be able to cope with such an increasing cost. The shortage of medical staffs including doctors, nurses, and medical assistants is another important issue. The production of the medical staffs also calls for an enormous amount of budget and a long period before they can be in reliable service. It is widely recognized that the utilization of appointment system is a systematic management of outpatient service.

The appointment system is expectedly to reduce the arrival variability and to increase the convenience of patients. Design of appointment system is, therefore, of paramount importance in order to attain effective and efficient outpatient services under management constraints. The research on appointment system has continuously proliferated since the pioneering work from Bailey [Bailey 1952]. A comprehensive review of the literature on appointment system can be found in [Cayirli and Veral 2003]. The appointment rule and the patient information are two main factors that are considered in the analysis and design of appointment system. In view of system analysis and design, the appointment rule represents the appointment system and its operational mechanism whereas the patient information is the system inputs. The appointment rule is characterized by its parameters including block size, initial block, appointment interval, time begin session, and time end session. The patient information includes their punctuality, presence status, need of second consultation, etc. The performance of an appointment system can be measured in various manners. The basic performance measures consist of waiting time of patients, idle time of doctors, and overtime of doctors.

The analysis of the performance of appointment system has received perpetual attention since the pioneer work from Bailey. Different appointment rules and their performances have been studied, compared, and reported under various situations of outpatient services. Both analytical and simulation-based approaches are employed in the analysis [Cayirli and Veral 2003]. The analysis of appointment system is aimed at the evaluation of the system performance. Prospect appointment systems can be obtained from the comparative analysis of different appointment systems. Although the comparative analysis can inform which appointment system has best performances, it may not reveal the systems that meet desirable performances. This depends heavily on whether or not the systems and their associated parameters that result in desirable performances are included in the comparative analysis. In other words, the comparative analysis may not recognize and thus overlook the systems and their associated parameters that render desirable performances. Such a drawback also occurs in the case that the decision on the type of appointment system has already been made, i.e. fixed appointment rule, but only the selection of its parameters need to

be carried out under the condition of producing desirable performances. Search for the appointment systems, i.e. the appointment rules and parameters that lead to desirable performances is thus another crucial issue. The search of such prospect systems can be carried out in terms of an optimization problem and is referred to as the optimal design of appointment system.

Various techniques have been employed in the optimal design of appointment system. [Fries and Marathe 1981] use dynamic programming to determine the optimal block sizes for the next period given that the number of patients remaining to be assigned is known. [Liao et al. 1993] apply dynamic programming to determine the optimal block sizes when service times are Erlang. [Liu and Liu 1998] develop a dynamic programming formulation to find the optimal block sizes in their study on a queuing system with multiple doctors with random arrival times. [Pegden and Rosenshine 1990] employ a Markov-chain based procedure to compute the optimal appointment intervals. [Robinson and Chen 2001] formulate the problem of finding the optimal appointment times as a stochastic linear program and solve it using Monte-Carlo integration. [Denton and Gupta 2001] present a two-stage stochastic linear programming model to determine the optimal appointment intervals. [Vanden Bosch et al. 1999] propose a fathoming approach to solve the same problem as [Liao et al. 1993]. [Kaandorop and Koole 2007] introduce a local search procedure to determine the optimal schedule with a weighted average of expected waiting times of patients, idle time of the doctor and tardiness as objective. While the afore-mentioned optimization techniques have proved the capability in obtaining solutions, their applicability is nevertheless limited to specific types of problem. In addition, certain problem characteristics need to be fulfilled in order to guarantee the access of global optima, e.g. the so-called multimodularity in [Kaandorop and Koole 2007].

The appointment system to be considered in this resesearch belongs to the class of individual block/fixed interval [Cayirli and Veral 2003]. One of the common characteristics of almost all literatures on outpatient scheduling is the limitation of the model to a single doctor with one queue. There are some studies that consider a parallel number of doctors. The analysis study from [Fetter and Thompson 1966]

considers three doctors. [Liu and Liu 1998] include two to five doctors in their simulation study. The present study takes into account a parallel number of doctors, i.e. multiple doctors, to make the problem become more realistic. However, only appointed patients will be exclusively considered, i.e. excluding walk-in patients. Accordingly, the disturbance against the appointment times from such a flow of walk-in patients is eliminated. The patient punctuality is defined in terms of the difference between the time of appointment and the time of patient arrival, which could be result in either lateness or earliness. The arrival times of the appointed patients can be random around the appointed times. Regarding the presence status of an appointed patient, it is possible that that the patient will not actually show up for the appointment, i.e. no-show patient. As no-show cases are unavoidable, it is also incorporated into the model of the appointment system. In addition, some patients require second consultation to the same doctors after their laboratory tests. The fact of multiple doctors, no-show patients, and second consultation is simultaneously treated with the objective to emulate real service situations. The optimal design is formulated in form of a constrained optimization problem where the number of doctors and the appointment interval are two design variables, given an expected number of appointed patients. This type of problem can be viewed as a resource planning and is critical to the management subject to the scarcity of doctor and the limitation of financial budget. The problem aspect applies to the planning of new outpatient departments or the improvement of existing ones in which the human resource and supporting budget is a major concern. The number of doctors and the appointment interval need to be determined in such a way that the service can accommodate outpatients at desirable degrees of performance and constraint satisfaction. It should be noted that the problem aspect represent one in other possible alternatives for efficient and effective planning or improvement of outpatient service.

The following definitions, characteristics, and rules are applied to the appointment system of interest.

1. The appointment system is in the class of parallel individual block/fixed interval.

2. There are n_D parallel doctors. It is assumed that there is no difference in the capability of all doctors. For a given number of patients $N_{patient}$, the patients will be distributed to each doctor as uniformly as possible. For example, when n_D is 3 and $N_{patient}$ is 18, the number of patients assigned to each doctor is equal to 6. However, when n_D is 3 and $N_{patient}$ is 20, two doctors take care of 7 patients and one doctor is responsible for 6 patients. The number of patients assigned to the jth doctor is denoted as NO_j.

3. It is possible that an appointed patient may be absent. The absence probability of each patient is equally designated to p_{abs}.

4. It is possible that an appointed patient can have laboratory tests. Each patient has an equal probability of having laboratory tests of p_{lab}. After the tests, that patient needs another consultation. The second consultation will be considered as another additional appointment case under the same doctor. Accordingly, a patient who has laboratory tests creates two consultation cases. If the patient that needs the second consultation arrives at the doctor room before the appointed patient in the schedule, the second-time patient will be given a priority to see the doctor. Otherwise, the second-time patient can see the doctor after the appointed patient in the schedule has finished the consultation. In other words, the First-Come-First-Serve (FCFS) principle is used when there is the interruption in the original schedule from second-consultation patients. It is reminded that all present patients have at least one consultation, i.e. their first consultation cases.

5. Each doctor may have different number of appointment blocks. Each ith block consists of only one consultation case. NP_j is the total number of consultation cases, including both first and second consultations, under the jth doctor. This number counts each second consultation as a case of treatment. Consequently,.

$$NP_j \geq NO_j$$

6. The appointment interval is denoted by Δt_{block}. The appointment time at the beginning of the ith appointment block is designated to t_i. Without loss of generality, t_1 is set equal to zero.

7. The office hour is ended at t_f.

The graphical representation of an appointment system is shown in Figures 2.1 and 2.2. Figure 2.1 shows an appointment system belonging to a doctor. The

system in Figure 2.1(a) does not include the no-show and second-consultation patients. Figure 2.1(b) depicts an appointment system with a no-show patient. Figure 2.1(c) includes both no-show and second-consultation patients. It should be noted that the original ranks of appointed patients has been changed due to the interference from the patients requiring second consultations. The ranks of the patients subjected to the change are expressed in a format of two numbers. For example, 13(12) means that this patient is originally at the 12^{th} rank in the queue but is then shifted to the 13^{th} rank because of the interference. Figure 2.2 represents an appointment system with multiple doctors, no-show patients, and patients requiring second consultations.

The variables and terms that are related to the service and performance indices will be now defined and described.

A_{ij} : arrival time of the *i*th-block of consultation case (either first or second) under the *j*th doctor in the appointment system

L_{ij} : length of service time for the *i*th-block of consultation case (either first or second) under the *j*th doctor in the appointment system

B_{ij} : starting service time of the *i*th-block of consultation case (either first or second) under the *j*th doctor in the appointment system

E_{ij} : ending service time of the *i*th-block of consultation case (either first or second) under the *j*th doctor in the appointment system

The arrival time of each consultation case is classified into two types. For the first consultation case, the arrival time is related to the appointment time as follows:

$$A_{ij} = t_i + \Delta_{ij} \qquad (2.26)$$

where Δ_{ij} is the time deviating from the appointment time t_i. The deviation time can be random and thus treated as a random variable. The punctuality of an appointed patient is interpreted from the condition

$$\Delta_{ij}\begin{cases}<0 & \text{early arrival}\\ =0 & \text{punctual arrival}\\ >0 & \text{late arrival}\end{cases} \qquad (2.27)$$

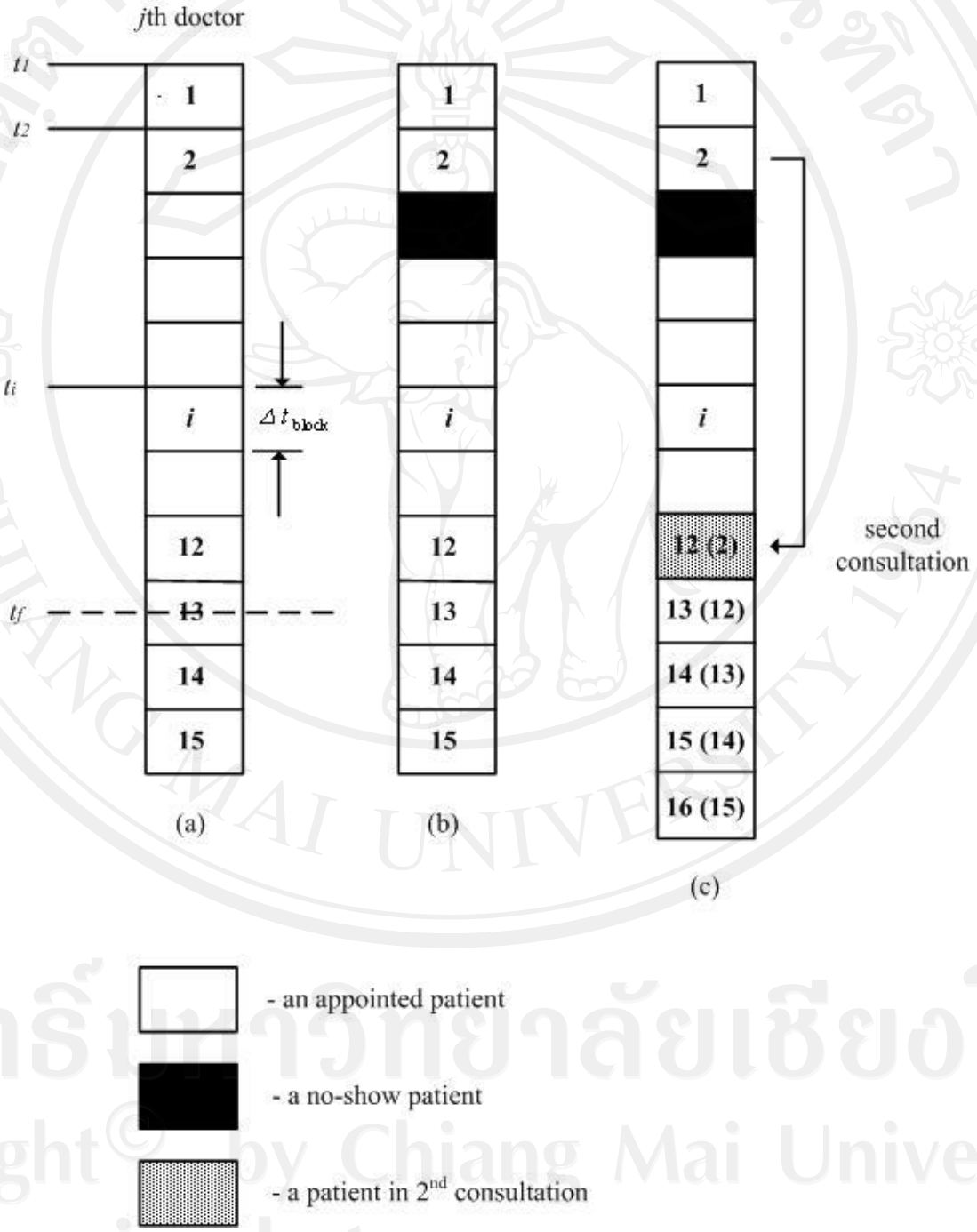

Figure 2.1. A part of an appointment system for a doctor.

When considering that the earliness or waiting prior to appointment time is not a consequence of the appointment system as in [Cayirli and Veral 2003], then Δ_{ij} is defined as

$$\Delta_{ij}\begin{cases}=0 & \text{early and punctual arrival}\\ >0 & \text{late arrival}\end{cases} \qquad (2.28)$$

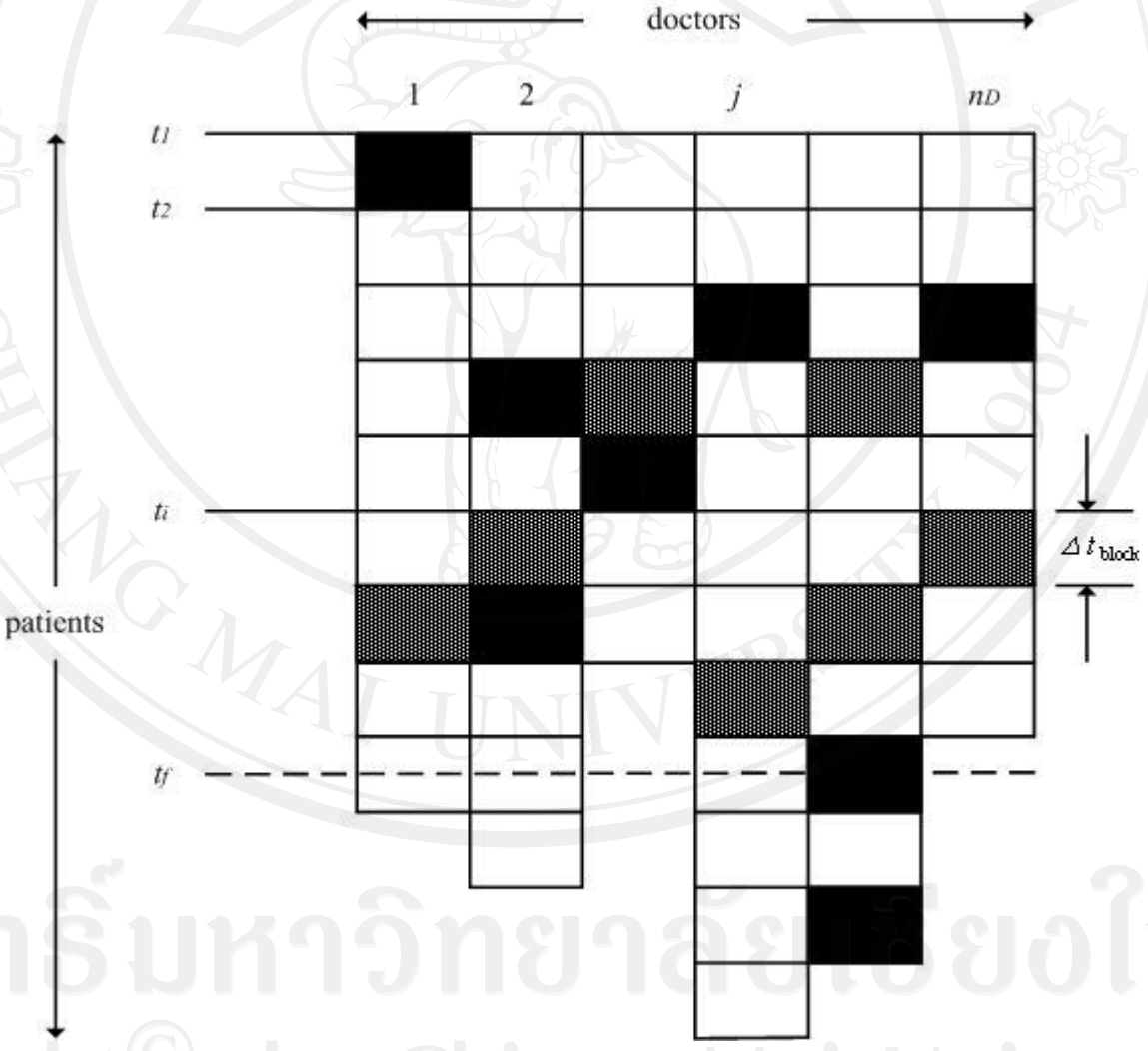

Figure 2.2. An appointment system with multiple doctors, no-show patients, and patients

requiring second consultations.

For the second consultation case, the arrival time is given by

$$A_{ij} = EF_{ij} + TL_{ij} \tag{2.29}$$

where EF_{ij} is the ending service time after the first consultation and TL_{ij} is the time required for the laboratory tests of that patient, respectively. The ending service time after the first consultation can be computed from Eq. (2.31).

The starting service time B_{ij} is obtained from

$$B_{ij} = max\left(A_{ij}, E_{(i-1)j}\right) \quad ; \ i = 2,..., NP_j \tag{2.30.1}$$

and

$$B_{1j} = max\left(A_{1j}, t_1\right) \tag{2.30.2}$$

which reflects the fact that the first patient to each doctor can have the healthcare service only after the starting office hour.

The ending service time of each consultation case, i.e. first or second consultation, is defined as

$$E_{ij} = B_{ij} + L_{ij} \tag{2.31}$$

L_{ij} is equal to zero if the *i*th-block patient under the *j*th doctor is absent or no-show. In addition, when the patient under consideration requires the second consultation and E_{ij} corresponds to the ending service time after the first consultation, then E_{ij} is further used as EF_{ij} for the computation of the arrival time for the corresponding second consultation case. That is

$$EF_{ij} = E_{ij} \tag{2.31}$$

for its used in Eq. (2.28). It should be noted that the length of service time L_{ij} is separated into two cases in all mathematical expressions. In the first consultation case, the length of service time for the first consultation $L1_{ij}$ must be used for L_{ij}, i.e. setting

$$L_{ij} = L1_{ij} \tag{2.33.1}$$

The second consultation case fixes the length of service time for the second consultation $L2_{ij}$ for L_{ij}

$$L_{ij} = L2_{ij} \tag{2.33.2}$$

Next the relevant performance indices will be defined. First, the waiting time W_{ij} of the ith-block of consultation case (either first or second) under the jth doctor is

$$W_{ij} = max\left(0, B_{ij} - A_{ij}\right) \tag{2.34}$$

The total waiting time corresponding to the service from the jth doctor W_j is

$$W_j = \sum_{i=1}^{NP_j} W_{ij} \tag{2.35}$$

The total waiting time in the appointment system W_T is thus

$$W_T = \sum_{j=1}^{nD} W_j \tag{2.36}$$

The average waiting time of a patient W_A is

$$W_A = \frac{1}{N_P n_D} W_T \tag{2.37}$$

where

$$N_P = \sum_{j=1}^{nD} NP_j \tag{2.38}$$

The overtime of the jth doctor OT_j is obtained from

$$OT_j = max\left(0, E_{NP_j j} - t_f\right) \tag{2.39}$$

where $E_{NP,j}$ is the ending service time of the last consultation case under the jth doctor. The definition of the jth-doctor overtime implies that there is no overtime if the doctor finishes the work before the office hour.

The total overtime in the appointment system OT_T is

$$OT_T = \sum_{j=1}^{nD} OT_j \tag{2.40}$$

The average overtime for a doctor OT_A is

$$OT_A = \frac{1}{n_D} OT_T \tag{2.41}$$

The *j*th doctor idle time incurred just before the arrival of the *i*th-block of consultation case (either first or second) is

$$IT_{ij} = max\left(0, A_{ij} - E_{(i-1)j}\right) \quad ; \ i = 2,..., NP_j \tag{2.42.1}$$

and

$$IT_{1j} = max\left(0, A_{1j} - t_1\right) \tag{2.42.2}$$

The total idle time of the *j*th doctor IT_j is

$$IT_j = \begin{cases} \sum_{i=1}^{NP_j} IT_{ij} & ; OT_j \geq 0 \\ \sum_{i=1}^{NP_j} IT_{ij} + \left|OT_j\right| & ; OT_j < 0 \end{cases} \tag{2.43}$$

The inclusion of the overtime term into the computation of the total idle time suggests that the free time of the doctor before the end of the office hour be considered as an idle time as well.

The total idle time in the appointment system IT_T is

$$IT_T = \sum_{j=1}^{nD} IT_j \tag{2.44}$$

The average idle time for a doctor IT_A is

$$IT_A = \frac{1}{n_D} IT_T \tag{2.45}$$

It should be noted that the performance indices as defined above can be combined in a various ways to establish the performance functions of the appointment system. As an example, the performance of an appointment system is measured through the expected total cost of appointment system $E[C_T]$ as defined by

$$E[C_T] = c_W E[W_T] + c_{OT} E[OT_T] + c_{IT} E[IT_T] \tag{2.46}$$

where c_W, c_{OT}, and c_{IT} is the cost per time unit associated to W_T, OT_T, and IT_T, respectively. The symbol $E[f(\boldsymbol{X})]$ denotes the expectation of a function $f(\boldsymbol{X})$. A comprehensive collection of performance measures used in the literature can be found in [Cayirli and Veral 2003].

As mentioned in the previous sections, a number of appointed patients may not appear at the times of appointment and some appointed patients need to have laboratory tests and thus their second consultation. In addition, the patient punctuality, i.e. the deviation time from the appointment time, can be random. All of these incidents can be rationally modeled as uncertain events. The framework of probability theory will be utilized in this study to model and measure the uncertainty. The inclusion of these random events makes the defined performance indices become uncertainty too. The measurements of the uncertain performance indices will be then carried out using probabilistic measures. Consequently, the measurements in the system performance through the performance functions will be done in turn using the probabilistic measures as well.

The optimal design of this appointment system is defined as follows:

$$\min_{(n_D, \Delta t_{\text{block}})} \; E[C_T(n_D, \Delta t_{\text{block}})]_{Npatient=50} = c_W E[W_T(n_D, \Delta t_{\text{block}})]_{Npatient=50} + c_{OT} E[OT_T(n_D, \Delta t_{\text{block}})]_{Npatient=50} + c_{IT} E[IT_T(n_D, \Delta t_{\text{block}})]_{Npatient=50} \tag{2.47}$$

Subject to

$$1 \le n_D \le 50 \tag{2.48}$$

$$\Delta t_{\text{block}} > 0 \tag{2.49}$$

$$E\left[W_A\left(n_D, \Delta t_{\text{block}}\right)\right]_{Npatient=50} \le \delta_W \quad (2.50)$$

$$E\left[OT_A\left(n_D, \Delta t_{\text{block}}\right)\right]_{Npatient=50} \le \delta_{OT} \quad (2.51)$$

$$E\left[IT_A\left(n_D, \Delta t_{\text{block}}\right)\right]_{Npatient=50} \le \delta_{IT} \quad (2.52)$$

where δ_W, δ_{OT}, and δ_{IT} are the thresholds of the average waiting time of a patient, the average overtime for a doctor, and the average idle time for a doctor, respectively.

2.2 SOLUTION METHODOLOGY

GA is a stochastic search technique based on the mechanism of natural selection. It combines Darwin's principle of survival of the fitter and a structured information exchange using randomized operators to evolve an efficient search mechanism. GA is naturally suitable to combinatorial optimization problems. Since the first problem considered here is a combinatorial optimization problem in which the combination of $\delta_{Aj}(t_i)$'s that leads to the optimal solution needs to be determined, GA is selected as a tool for searching the optimal values of $\delta_{Aj}(t_i)$'s. A binary coding for real value will be used to represent $\delta_{Aj}(t_i)$. The combination of these strings forms a chromosome.

The fitness function of a chromosome $F(\Delta)$ is defined as

$$F(\Delta) = \frac{1}{O(\Delta)} \quad ; \Delta = \{\delta_{Aj}(t_i) |\ i = 1,\ldots,NYear \text{ and } j = 1,\ldots,NAge\} \quad (2.26)$$

, in which $O(\Delta)$ is defined as

$$O(\Delta) = \begin{cases} ERR(\Delta) & ; \Delta \text{ is feasible} \\ ERR(\Delta) + \sum_{j=1}^{NCon} c_k v_k(\Delta) & ; \Delta \text{ is infeasible} \end{cases} \quad (2.27)$$

$v_k(\Delta)$: The violation magnitude of the kth constraint

$<v_k(\Delta)>$: The average of $v_k(\Delta)$ over the population

c_k : The penalty parameter for the kth constraint defined at each generation

$NCon$: The total number of constraints

$F(\Delta)$: The fitness function

ε : The tolerance for the 100-percent criteria

As seen from (2.26) and (2.27), a penalty approach is employed to handle the constraints. The penalty approach is the well-known and widely applied technique for handling constraints in GA. In order to circumvent the difficulty in assigning the penalty factor c_k, the method of adaptive penalty is selected for this study. Different schemes in the method of adaptive penalty are proposed. More specifically, an adaptive penalty scheme which had been proposed by [Barbosa and Lemonge 2003] and was then modified by [Obadage and Harnpornchai 2006] is utilized herein. Such an adaptive penalty scheme has an advantage of its algorithmic simplicity. Accordingly, the penalty factor c_k is given by

$$c_k = \left|\max\left(ERR_{\inf}(\Delta)\right)\right| \frac{< v_k(\Delta) >}{\sum_{l=1}^{NCon} \left[< v_l(\Delta) >\right]^2} \tag{2.28}$$

Note that the equality constraint (2.8) is modified to be an inequality constraint

$$\frac{\left|\sum_{j=1}^{NAge} E_{Aj}(t_i) - 100\right|}{100} \le \varepsilon \quad ; \forall i \tag{2.29}$$

,where $max(ERR_{inf}(\Delta))$ is the maximum of the objective function values for the current population in the infeasible region. The tolerance ε can be arbitrarily set but is normally a small value.

An adaptive penalty GA is introduced as a tool for the second optimization problem too. The design variables are encrypted using the binary-real coding scheme. According to the adaptive penalty GA used, the fitness function $F(\boldsymbol{S})$ is defined as

$$F(\boldsymbol{S}) = \begin{cases} 1/O(\boldsymbol{S},\boldsymbol{D}) & ;\ \boldsymbol{S} \text{ is feasible} \\ 1/\left[O(\boldsymbol{S},\boldsymbol{D}) - \sum_{j=1}^{NYear} k_j v_j(\boldsymbol{S},\boldsymbol{D})\right] & ;\ \boldsymbol{S} \text{ is infeasible} \end{cases} \tag{2.30}$$

The adaptive penalty scheme is given by

$$k_j = \left|max\left(O^{\inf}(\boldsymbol{S},\boldsymbol{D})\right)\right| \frac{< v_j(\boldsymbol{S},\boldsymbol{D}) >}{\sum_{l=1}^{NYear} \left[< v_l(\boldsymbol{S},\boldsymbol{D}) >\right]^2} \tag{2.31}$$

where $max(O^{\inf}(\boldsymbol{S},\boldsymbol{D}))$ is the maximum of the objective function values in the current population in the infeasible region, $v_j(\boldsymbol{S},\boldsymbol{D})$ is the violation magnitude of the jth constraint. $<v_j(\boldsymbol{S},\boldsymbol{D})>$ is the average of $v_j(\boldsymbol{S},\boldsymbol{D})$ over the current population. k_j is the penalty parameter for the jth constraint defined at each generation. The violation magnitude is defined as

$$v_j(\boldsymbol{S},\boldsymbol{D}) = \begin{cases} \left|TC_j(\boldsymbol{S},\boldsymbol{D}) - TBU_j\right| & ;\ TC_j(\boldsymbol{S},\boldsymbol{D}) - TBU_j > 0 \\ 0 \end{cases} \tag{2.32}$$

In the next chapter, the numerical examples corresponding to those two problems will be performed to clarify the proposed methodology.