JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

3/2002

Special issue edited by Janusz Granat

Editorial Board

Editorial Advisory Board

ISSN 1509-4553

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

The dynamics of telecommunication sector development and strong competition in this sector in most countries induce telecommunication operators to use new computerised tools for decision support. Such tools might help in planning network development, in assessing financial and economic standing, in network design and management, in corporate management including strategic decisions and negotiations of interconnection agreements. The intensive development and decreasing costs of information technology increase the availability of such tools. National Institute of Telecommunications organised *The First International Conference on Decision Support for Telecommunications and Information Society DSTIS-2001* in Warsaw (11th – 14th July 2001). The aim of this conference was to bring together international experts and researchers working in decision support and related fields, in order to present diverse methods and tools, which are applied in telecommunications industry and information society. This issue presents the most interesting scientific papers presented at this conference.

The first paper presents the concept of AmI (ambient intelligence) and AmI scenarios of ISTAG (Information Society Technology Advisory Group of European Commission). The requirements of intelligence versus decision support are then discussed. Resulting challenges for decision support systems (DSS) in telecommunications are then outlined and conclusions presented. The second paper outlines the basic concepts of the rough sets theory and a simple tutorial example, concerning churn modeling in telecommunications. Rough set theory deals with imperfect knowledge and might be useful in modeling various telecommunications problems. The next paper shows how to use the data mining techniques to improve knowledge about customers of the operators. One of the main issues of telecommunication operators today is to be able not only to store and manage the huge amount of data generated by applications, but also to give value to this data. Further, there is a paper devoted to stochastic analysis of telecommunication networks. It presents analytical properties of a stochastic teletraffic system with MMPP (Markov modulated Poisson process) input and an access function.

The next three papers focus on intelligent systems. However, they do not present direct examples of applications in telecommunication industry but we decide to include them in this issue because developments of intelligent systems for this industry are very promising.

There is a paper that presents effectiveness of active forgetting in machine learning. In many telecommunications problems the situation changes very often over time. In machine learning, therefore, decision rules are needed to adapt for such changeable situations. Additional learning should be made on new data. On the other hand, since rules for classification becomes more complex with only additional learning, appropriate forgetting is also necessary. The next paper addresses a problem in the area of intelligent knowledge-based systems. Knowledge is a central component in any intelligent knowledge-based system. Results presented in this paper obtained from different investigations indicate the potential of the approach based on fuzzy control systems for generation of knowledge. Further, there is a paper that shows the application of multi-agent utility theory for ethical conflict resolution. This paper shows how to construct a two-attribute group disutility function for two conflicting decision makers, taking into account the property of utility independence and/or convex dependence between them.

We have also two papers that show the application of optimization methods. Telecommunication networks are expected to satisfy the increasing demand for various services. Hence, it becomes critical to allocate network resources to provide high level performance of all services. We have paper that introduce and analyze a solution concept of the conditional minimax as a generalization of the minimax solution concept extended to take into account the number of services related to worst performances. Namely, for a specified portion of demand authors take into account the corresponding portion of the maximum results and they consider their average as the worst conditional mean to be minimised. The authors show that the minimisation of the worst conditional mean can be defined by linear objective and a number of auxiliary linear inequalities. The second paper that uses optimization techniques exploit the partially linear structure of the nonlinear multicommodity flow optimization problem. At the end we have two application papers. The first one focus on web cache management. The second one presents the analytical system for testing a telecommunication network.

Janusz Granat
Guest Editor

Paper

The concept of ambient intelligence and decision support for telecommunications

Andrzej P. Wierzbicki

Abstract — The paper presents first the concept of AmI (ambient intelligence) and AmI scenarios of ISTAG (Information Society Technology Advisory Group of European Commission). The requirements of intelligence versus decision support are then discussed. Resulting challenges for decision support systems (DSS) in telecommunications are then outlined and conclusions presented.

Keywords — ISTAG, ambient intelligence, decision support for telecommunications.

1. The concept of ambient intelligence

European Commission has committees and advisory groups related to the Framework Programme. Committees are composed of the delegates of member or participating countries, advisory groups consists of experts selected and nominated by the Commission. Information Society Technology Advisory Group (ISTAG) works since 1999 on the vision of IST development in the 5th and 6th Framework Programme. This vision is summarised by the concept of AmI.

Shortly, *ambient intelligence is a future information society environment with intelligence embedded anywhere but in an unobtrusive fashion*, with the emphasis on:

- greater user-friendliness;
- more efficient services support;
- user empowerment;
- support for human interactions.

In AmI environment, people are surrounded by intelligent intuitive interfaces embedded in all kind of objects; this environment is capable of recognising and responding to the presence of different individuals in a seamless, unobtrusive and often invisible way.

This concept can be illustrated by a “simple” example from housing telematics:

- imagine an ordinary room;
- a person coming and asking the room “Connect me to Maria”;
- a wall changing into a huge screen;
- a hidden personal communication interface, capable of:
  - recognising the coming person and guessing who is “Maria”;
  - making local broadband connection to the backbone network and searching for “Maria”, who might be travelling;
  - displaying video communication with diverse options.

In other words, AmI is a vision of next big generation of communication culture that relates to internet such as internet relates to classical voice telephony.

2. Ambient intelligence scenarios of ISTAG

In order to determine critical aspects of AmI, ISTAG realised that socio-economic demand is decisive for broad acceptance of new technologies. Thus, the relevant question is: what aspects of AmI would people soonest buy? In order to answer this question, ISTAG asked the Institute for Prospective Technology Studies (IPTS) in Seville to develop “scenarios for ambient intelligence 2010”. This date might seem optimistic for the implementation of actually a new paradigm of human communication, but IPTS developed [1] scenarios, 5 critical requirements, roadmaps, main research implications and opportunities, etc.; allthis was discussed and after corrections accepted by ISTAG. We shall shortly characterise these scenarios, together with comments on their possible realisation time. After all, IPTS did not realise that digital television was conceived 40 years ago and still is not broadly implemented.

Scenario “Maria” – road warrior. Maria is a travelling businesswomen, with only one personal communication device that helps her to organise everything – communicate broadly in business and with family, find data and files for business presentations, organise travel, find rental cars, organise business schedules, etc. The necessary technological requirements for this scenario include: a seamless and intelligent mobile-fixed broadband network, a novel personal communication device, etc. This might be realised in one or two decades, and the possible market demand would be immediately large after sufficient technological development.

Scenario "Dymitrios" – digital Me (DMe). Dymitrios is a personal communication device with sufficient intelligence to be a personal secretary simulating the actual person in diverse contacts of secondary importance, while recognising situations of prime importance and arranging actual contacts, including multilateral conferences etc. The necessary technological requirements include much higher demands on computer intelligence than in scenario "Maria" which implies that the scenario might possibly require longer, two to three decades, to be realised. Possible demand would also be not necessarily immediately large: there are social reservations to computers trying to outsmart people, known well to specialists in applications of decision support.

Scenario "Carmen" – traffic, sustainability, commerce. This scenario concerns ambient intelligence environment for travel and commerce. It assumes full automation of traffic, vehicle area networks, micro-payment systems for collecting fares, full integration with metro and other transportation networks as well as with goods distribution networks. Technological demands for this scenario are rather high: it demands full traffic and logistics automation. Thus, the time required to realise this scenario might be longest, possibly three or four decades. Possible demand would be immediately large, but after sufficient technological development.

Scenario "Annette and Solomon" – an ambient for social learning. This scenario concerns an ambient - an agent working as an automated mentor, together with local network environment combined with global resources for social learning, group dynamics, etc. The ambient combines distant education features with psychological educational aspects. Technological demands are not much higher than in the scenario "Dimitrios", hence the time for implementation might be two to three decades. Possible demand might be large, but there are various psychological and educational reservations, similarly as in scenario "Dimitrios".

Very interesting are critical factors for the implementation of the above scenarios, specified in the IPTS-ISTAG report. These factors are socio-political, business-economic, technological, etc. The main technological factors are listed as:

- very unobtrusive hardware;
- seamless mobile-fixed communication infrastructure;
- dynamic and massively distributed device networks;
- natural feeling human interfaces;
- dependability and security.

What is not listed in the critical technological factors, although obviously such a factor but overlooked by IPTS is:

- very fast development of computer intelligence and decision support methods, combined with telecommunication applications.

3. Computer intelligence versus decision support

Classical computer intelligence research is based on a natural objective that computer *should become* more intelligent than people. In opposition, applied decision support is based on the premise that computers *should serve people* (decision makers), in particular – *do not outsmart them*, because they otherwise *do not use* computer support.

The vision of AmI tries to overcome this dilemma by insisting that *computers and networks should become much more intelligent, but should nevertheless serve people.* It is decisive for future development and implementation of AmI whether we will be able to actually overcome this dilemma.

In research, the goal of making computers as intelligent as possible is fully legitimate. *In applications*, trying to outsmart computer users leads always to trouble.

For example, *impersonating* the actual person in the scenario "Dimitrios" might lead to serious trouble (since we might not know what our DMe has said on our behalf). In the scenario "Annette and Solomon", the ambient *pretending* to be the mentor might lead to psychological difficulties. It is well known in decision support history that when using optimisation techniques, researchers insisting that their user (e.g. designer) should be able to define a scalar objective function or utility function (that would actually *impersonate* the user) could not apply their techniques in practice. We can show many such examples.

Thus, the main challenge before us – specialists in computer intelligence and decision support – is to make computers *more intelligent but much more user-friendly than intelligent*.

4. Challenges for decision support in telecommunications

However, there is no doubt that the concept of AmI defines many new specific challenges for decision support and computer intelligence in telecommunications. It is not possible to list all such challenges, we shall show only some examples here. To these challenges belong:

Massive and diversified data processing. This includes data integration, warehouses, data models, data mining and knowledge extraction, all also in distributed network applications.

From data to sophisticated substantive models. Classical knowledge extraction discerning logical patterns from data relies only on a narrow definition of knowledge. In applied decision support, an important concept is that of *substantive models* – any type of models, be it in logical or analytical form, describing knowledge pertinent for given application. Substantive models are essential in engineering design, in business forecasting, etc. The development of such models is related to such subjects as the art and science of model building, the standards for computerised models, special platforms and languages for model building, etc.

Challenges corresponding to ISTAG scenarios. We can list many computer intelligence and decision support challenges related to IPTS-ISTAG scenarios. For example, when starting with Maria, we need the detection not only of geographical location but also of close information and transportation service centres. The list of such challenges is enormous; recently, ISTAG Working Group 9 tried to define and list most of challenges related to software technologies resulting from AmI vision [2].

Specific telecommunication issues. There are also many specific issues in telecommunications that relate to computer intelligence and decision support and will have a direct relation to AmI vision. We list below only some of them:

- intelligent mobile services in 3rd generation (UMTS); DSS promoting such services;
- DSS promoting the use and management of digital interactive television networks;
- DSS in quality control of telecommunication services;
- DSS in electromagnetic spectrum management;
- DSS in regulation of interconnection issues;
- DSS in network management;
- DSS in enhancing dependability and security of telecommunication services; etc.

5. Conclusions

We shall list shortly here main conclusions:

- IPTS-ISTAG AmI scenarios might be futuristic, but they help to reflect on future developments;
- we should make computers and networks much more intelligent, but even more user-friendly and serving people than intelligent;
- we should promote the transition from data-based DSS to more sophisticated substantive models in DSS and model-based DSS;
- there are many specific challenges for decision support and computer intelligence in telecommunications;
- future civilisation will depend on the way how we shall respond to these challenges.

References

[1] *Scenarios for Ambient Intelligence 2010*, http://www.cordis.lu/ist/istag.htm

[2] *Software Technologies, Embedded Systems and Distributed Systems. An European Strategy Towards an Ambient Intelligent Environment.* WG9 Report v2.2.

Andrzej Piotr Wierzbicki born June 29, 1937 in Warsaw. Graduated as Master of Engineering at the Faculty of Electronics, Warsaw University of Technology (WUT), in 1960. Ph.D. degree at this University in 1964, for a thesis on nonlinear feedback systems; D.Sc. degree in 1968, for a thesis on optimisation of dynamic systems. In 1971–75 a Deputy Director of the Institute of Automatic Control, later a Deputy Dean of the Faculty of Electronics, WUT. In 1975–78 the Dean of the Faculty of Electronics WUT. Since 1978 worked with the International Institute for Applied Systems Analysis in Laxenburg n. Vienna, Austria; 1979–84 as the chairman of the theoretical branch, Systems and Decision Sciences Program, of this Institute. From 1985 back in the Institute of Automatic Control, WUT, as a Professor of optimisation and decision theory. In 1986–91 scientific secretary, currently member of presidium of the Committee of Future Studies “Poland 2000” (in 1990 renamed “Poland in XXI Century”) of P.Ac.Sc. In 1991 elected a member of the State Committee for Scientific Research of Republic of Poland and the chairman of its Commission of Applied Research; contributed to basic reforms of Polish scientific system in 1991–94. Deputy chairman of the Council of Polish Foundation for Science in 1991–94, chairman of scientific councils of NASK (National Scientific and Academic Computer Network in Poland) and PIAP (the Industrial Institute of Measurements and Control). In 1991–96 the editor in chief of the quarterly “Archives of Control Sciences” of P.Ac.SC. In 1992 received (as first European researcher) the George Cantor Award of the International Society of Multiple Criteria Decision Making for his contributions to the theory of multiple criteria optimisation and decision support. Since 1996 the General Director of the National Institute of Telecommunications in Poland. In 2000 nominated as a member of the ISTAG (Information Society Technology Advisory Group) at European Commission. Since 2001 chairman of Advisory Group on Scientific International Cooparation of the State Committee for Scientific Research of Poland. Beside lecturing for over 40 years and promoting more than 80 master’s theses at WUT (Warsaw University of Technology), he also lectured at the Department of Mathematics, Information

Science and Mechanical Engineering of Warsaw University and in doctoral studies: at WUT, the Academy of Mining and Metallurgy, at the University of Minnesota, at the Illinois Technical University, Hagen University, and at the University of Kyoto. He also promoted 18 completed doctoral dissertations. Author of over 180 publications, including 11 books (4 monographs, 7 – editorship or co-authorship of international joint publications, over 50 articles in scientific journals (over 30 in international), 80 papers at conferences (68 at international, including over 48 published as chapters in books). He also authored 3 patents granted and applied industrially. Current interests include parallelisation of optimisation algorithms using multiple criteria approaches, diverse aspects of negotiation and decision support, including e.g. applications of fuzzy set theory for describing uncertainty in decision support models, multimedia software in computer networks, telematics in education, diverse issues of information society and civilisation. Languages: English, German, Russian (each fluent, beside native Polish). Member of IEEE,ISMCDM (International Society of Multiple Criteria Decision Making), SEP (Polish Society of Electrical Engineers), PTM (Polish Mathematical Society), PSKR (Polish Association for the Club of Rome).
e-mail: A.Wierzbicki@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paper

Rough set theory and its applications

Zdzisław Pawlak

Abstract — In this paper rudiments of the theory will be outlined, and basic concepts of the theory will be illustrated by a simple tutorial example, concerning churn modeling in telecommunications. Real life applications require more advanced extensions of the theory but we will not discuss these extensions here. Rough set theory has an overlap with many other theories dealing with imperfect knowledge, e.g., evidence theory, fuzzy sets, Bayesian inference and others. Nevertheless, the theory can be regarded as an independent, complementary, not competing, discipline in its own rights.

Keywords — rough set, decision rules, churn modeling.

1. Introduction

Rough set theory can be regarded as a new mathematical tool for imperfect data analysis. The theory has found applications in many domains, such as decision support, engineering, environment, banking, medicine and others.

This paper presents basis of the theory which will be illustrated by a simple example of churn modeling in telecommunications.

Rough set philosophy is founded on the assumption that with every object of the universe of discourse some information (data, knowledge) is associated. Objects characterized by the same information are *indiscernible (similar)* in view of the available information about them. The *indiscernibility relation* generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible (similar) objects is called an *elementary set*, and forms a basic *granule (atom)* of knowledge about the universe. Any union of some elementary sets is referred to as a *crisp (precise)* set – otherwise the set is *rough (imprecise, vague)*. Each rough set has boundary-line cases, i.e., objects which cannot be with certainty classified, by employing the available knowledge, as members of the set or its complement. Obviously rough sets, in contrast to precise sets, cannot be characterized in terms of information about their elements. With any rough set a pair of precise sets, called the *lower* and the *upper approximation* of the rough set, is associated. The lower approximation consists of all objects which *surely* belong to the set and the upper approximation contains all objects which *possibly* belong to the set. The difference between the upper and the lower approximation constitutes the *boundary region* of the rough set. Approximations are fundamental concepts of rough set theory.

Rough set based data analysis starts from a data table called a *decision table*, columns of which are labeled by *attributes*, rows – by *objects* of interest and entries of the table are *attribute values*. Attributes of the decision table are divided into two disjoint groups called *condition* and *decision* attributes, respectively. Each row of a decision table induces a *decision rule*, which specifies decision (action, results, outcome, etc.) if some conditions are satisfied. If a decision rule uniquely determines decision in terms of conditions – the decision rule is *certain*. Otherwise the decision rule is *uncertain*. Decision rules are closely connected with approximations. Roughly speaking, certain decision rules describe lower approximation of decisions in terms of conditions, whereas uncertain decision rules refer to the boundary region of decisions.

With every decision rule two conditional probabilities, called the *certainty* and the *coverage* coefficient, are associated. The certainty coefficient expresses the conditional probability that an object belongs to the decision class specified by the decision rule, given it satisfies conditions of the rule. The coverage coefficient gives the conditional probability of reasons for a given decision.

It turns out that the certainty and coverage coefficients satisfy Bayes' theorem. That gives a new look into the interpretation of Bayes' theorem, and offers a new method data to draw conclusions from data.

In the paper rudiments of the theory will be outlined, and basic concepts of the theory will be illustrated by a simple tutorial example of churn modeling. Real life applications require more advanced extensions of the theory but we will not discuss these extensions in this paper.

Rough set theory has an overlap with many other theories dealing with imperfect knowledge, e.g., evidence theory, fuzzy sets, Bayesian inference and others. Nevertheless, the theory can be regarded as an independent, complementary – not competing discipline, in its own rights.

More information about rough sets and their applications can be found in the references and the Web.

2. Illustrative example

Let us start our considerations from a very simple tutorial example concerning churn modeling in telecommunications, which is a simplified version of an example given in [1]. In Table 1, six facts concerning six client segments are presented.

In the table condition attributes describing client profile are: *In* – incoming calls, *Out* – outgoing calls within the same operator, *Change* – outgoing calls to other mobile operator, the decision attribute describing the consequence is *Churn* and *N* is the number of similar cases.

Each row in the table determine a decision rule. E.g., row 2 determines the following decision rule: *"if the number of incoming calls is high and the number of outgoing calls is high and the number of outgoing calls to the mobile operator is low then these is no churn"*.

According to [1]: *"One of the main problem that have to be solved by marketing departments of wireless operators is to find the way of convincing current clients that they continue to use the services. In solving this problems can help churn modeling. Churn model in telecommunications industry predicts customers who are going to leave the current operator"*.

Table 1
Client segments

Segment	*In*	*Out*	*Change*	*Churn*	*N*
1	medium	medium	low	no	200
2	high	high	low	no	100
3	low	low	low	no	300
4	low	low	high	yes	150
5	medium	medium	low	yes	220
6	medium	low	low	yes	30

In other words we want to explain churn in terms of clients profile, i.e., to describe market segments {4, 5, 6} (or {1, 2, 3}) in terms of condition attributes *In, Out* and *Change*.

The problem cannot be solved uniquely because the data set is *inconsistent*, i.e., segments 1 and 5 have the same profile but different consequences.

Let us observe that:

- segments 2 and 3 (4 and 6) can be classified as sets of clients who *certainly* do not churn (churn),
- segments 1, 2, 3 and 5 (1, 4, 5 and 6) can be classified as sets of clients who *possibly* do not churn (churn),
- segments 1 and 5 are *undecidable* sets of clients.

This leads us to the following notions:

- the set {2,3} ({4,6}) is the *lower approximation* of the set {1,2,3}({4,5,6}),
- the set {1,2,3,5} ({1,4,5,6}) is *the lower approximation* of the set {1,2,3} ({4,5,6}),
- the set {1,5} is the *boundary region* of the set {1,2,3}({4,5,6}),

which will be discussed in the next paragraph more exactly.

3. Information systems and approximations

In this section we will examine approximations more exactly. First we define a data set, called an information system.

An *information system* is a pair $S = (U,A)$, where U and A, are finite, nonempty sets called the *universe*, and the set of *attributes*, respectively. With every attribute $a \in A$ we associate a set V_a, of its *values*, called the *domain* of a. Any subset B of A determines a binary relation $I(B)$ on U, which will be called an *indiscernibility relation*, and defined as follows: $(x,y) \in I(B)$ if and only if $a(x) = a(y)$ for every $a \in A$, where $a(x)$ denotes the value of attribute a for element x. Obviously $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e., a partition determined by B, will be denoted by $U/I(B)$, or simply by U/B; an equivalence class of $I(B)$, i.e., block of the partition U/B, containing x will be denoted by $B(x)$. If (x,y) belongs to $I(B)$ we will say that x and y are *B-indiscernible (indiscernible with respect to B)*. Equivalence classes of the relation $I(B)$ (or blocks of the partition U/B) are referred to as *B-elementary sets* or *B-granules*.

Suppose we are given an information system $S = (U,A)$, $X \subseteq U$, and $B \subseteq A$. Let us define two operations assigning to every $X \subseteq U$ two sets $B_*(X)$ and $B^*(X)$, called the *B-lower* and the *B-upper approximation* of X, respectively, and defined as follows:

$$B_*(X) = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\},$$

$$B^*(X) = \bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \emptyset\}.$$

Hence, the *B-lower approximation* of a set is the union of all *B-granules* that are included in the set, whereas the *B-upper approximation* of a set is the union of all *B-granules* that have a nonempty intersection with the set. The set

$$BN_B(X) = B^*(X) - B_*(X)$$

will be referred to as the *B-boundary region* of X.

If the boundary region of X is the empty set, i.e., $BN_B(X) = \emptyset$, then X is *crisp (exact)* with respect to B; in the opposite case, i.e., if $BN_B(X) \neq \emptyset$, X is referred to as *rough (inexact)* with respect to B.

Thus, the set of elements is rough (inexact) if it cannot be defined in terms of the data, i.e. it has some elements that can be classified neither as member of the set nor its complement in view of the data.

4. Decision tables and decision rules

If we distinguish in an information system two disjoint classes of attributes, called *condition* and *decision attributes*, respectively, then the system will be called a *decision table* and will be denoted by $S = (U,C,D)$, where C and D are disjoint sets of condition and decision attributes, respectively.

Let $S = (U,C,D)$ be a decision table. Every $x \in U$ determines a sequence $c_1(x), \ldots, c_n(x)$, $d_1(x), \ldots, d_m(x)$, where $\{c_1, \ldots, c_n\} = C$ and $\{d_1, \ldots, d_m\} = D$.

The sequence will be called a *decision rule induced by x* (in S) and will be denoted by $c_1(x),\ldots,c_n(x) \rightarrow d_1(x),\ldots,d_m(x)$ or in short $C \rightarrow_x D$.

The number $supp_x(C,D) = |A(x)| = |C(x) \cap D(x)|$ will be called a *support* of the decision rule $C \rightarrow_x D$ and the number

$$\sigma_x(C,D) = \frac{supp_x(C,D)}{|U|},$$

will be referred to as the *strength* of the decision rule $C \rightarrow_x D$, where $|X|$ denotes the cardinality of X.

With every decision rule $C \rightarrow_x D$ we associate the *certainty factor* of the decision rule, denoted $cer_x(C,D)$ and defined as follows:

$$cer_x(C,D) = \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{supp_x(C,D)}{|C(x)|} = \frac{\sigma_x(C,D)}{\pi\big(C(x)\big)},$$

where $\pi\big(C(x)\big) = \frac{|C(x)|}{|U|}$.

The certainty factor may be interpreted as a conditional probability that y belongs to $D(x)$ given y belongs to $C(x)$, symbolically $\pi_x(D|C)$.

If $cer_x(C,D) = 1$, then $C \rightarrow_x D$ will be called a *certain decision* rule; if $0 < cer_x(C,D) < 1$ the decision rule will be referred to as an *uncertain decision rule*.

Besides, we will also use a *coverage factor* of the decision rule, denoted $cov_x(C,D)$ and defined as

$$cov_x(C,D) = \frac{|C(x) \cap D(x)|}{|D(x)|} = \frac{supp_x(C,D)}{|D(x)|} = \frac{\sigma_x(C,D)}{\pi\big(D(x)\big)},$$

where $\pi\big(C(x)\big) = \frac{|D(x)|}{|U|}$.

Similarly

$$cov_x(C,D) = \pi_x(C|D).$$

If $C \rightarrow_x D$ is a decision rule then $D \rightarrow_x C$ will be called an *inverse decision rule*. The inverse decision rules can be used to give *explanations (reasons)* for a decision.

For Table 1 we have the certainty and coverage factors are as shown in Table 2.

Table 2
Parameters of the decision rules

Decision rule	Strength	Certainty	Coverage
1	0.20	0.48	0.33
2	0.10	1.00	0.17
3	0.30	1.00	0.50
4	0.15	1.00	0.38
5	0.22	0.52	0.55
6	0.03	1.00	0.07

Let us observe that if $C \rightarrow_x D$ is a decision rule then

$$\bigcup_{y \in D(x)} \{C(y) : C(y) \subseteq D(x)\}$$

is the lower approximation of the decision class $D(x)$, by condition classes $C(y)$, whereas the set

$$\bigcup_{y \in D(x)} \{C(y) : C(y) \cap D(x) \neq \emptyset\}$$

is the upper approximation of the decision class by condition classes $C(y)$.

Approximations and decision rules are two different methods to express properties of data. Approximations suit better to express topological properties of data, whereas decision rules describe in a simple way hidden patterns in data.

5. Probabilistic properties of decision tables

Decision tables (and decision algorithms) have important probabilistic properties which are discussed next.

Let $C \rightarrow_x D$ be a decision rule and let $\Gamma = C(x)$ and $\Delta = D(x)$. Then the following properties are valid:

$$\sum_{y \in \Gamma} cer_y(C,D) = 1, \tag{1}$$

$$\sum_{y \in \Delta} cov_y(C,D) = 1, \tag{2}$$

$$\pi\big(D(x)\big) = \sum_{y \in \Gamma} cer_y(C,D) \cdot \pi\big(C(y)\big) = = \sum_{y \in \Gamma} \sigma_y(C,D), \tag{3}$$

$$\pi\big(C(x)\big) = \sum_{y \in \Delta} cov_y(C,D) \cdot \pi\big(D(y)\big) = = \sum_{y \in \Delta} \sigma_y(C,D), \tag{4}$$

$$cer_x(C,D) = \frac{cov_x(C,D) \cdot \pi\big(D(x)\big)}{\sum_{y \in \Delta} cov_y(C,D) \cdot \pi\big(D(y)\big)} = \frac{\sigma_x(C,D)}{\pi\big(C(x)\big)}, \tag{5}$$

$$cov_x(C,D) = \frac{cer_x(C,D) \cdot \pi\big(C(x)\big)}{\sum_{y \in \Gamma} cer_y(C,D) \cdot \pi\big(C(y)\big)} = \frac{\sigma_x(C,D)}{\pi\big(D(x)\big)}. \tag{6}$$

That is, any decision table satisfies Eqs.(1)–(6). Observe that formulae (3) and (4) refer to the well known *total probability theorem*, whereas (5) and (6) refer to *Bayes' theorem*.

Thus in order to compute the certainty and coverage factors of decision rules according to formula (5) and (6) it is enough to know the strength (support) of all decision rules only. The strength of decision rules can be computed from data or can be a subjective assessment.

6. Decision algorithm

Any decision table induces a set of *"if ... then"* decision rules.

Any set of mutually, exclusive and exhaustive decision rules, that covers all facts in S and preserves the indiscernibility relation included by S will be called a decision algorithm in S.

An example of decision algorithm in the decision Table 1 is given below:

		cer.
1)	*if (In, high) then (Churn, no)*	1.00
2)	*if (In, low) and (Change, low) then (Churn, no)*	1.00
3)	*if (In, med.) and (Out, med.) then (Churn, no)*	0.48
4)	*if (Change, high) then (Churn, yes)*	1.00
5)	*if (In, med.) and (Out, low) then (Churn, yes)*	1.00
6)	*if (In, med.) and (Out, med.) then (Churn, yes)*	0.52

Finding a minimal decision algorithm associated with a given decision table is rather complex. Many methods have been proposed to solve this problem, but we will not consider this problem here.

If we are interested in *explanation* of decisions in terms of conditions we need an *inverse* decision algorithm which is obtained by replacing mutually conditions and decisions in every decision rule in the decision algorithm.

For example, the following inverse decision algorithm can be understood as explanation of churn (no churn) in terms of client profile:

		cer.
1')	*if (Churn, no) then (In, high) and (Out, med.)*	0.33
2')	*if (Churn, no) then (In, high)*	0.17
3')	*if (Churn, no) then (In, low) and (Change, low)*	0.50
4')	*if (Churn, yes) then (Change, yes)*	0.38
5')	*if (Churn, yes) then (In, med.) and (Out, med.)*	0.55
6')	*if (Churn, yes) then (In, med.) and (Out, low)*	0.07

Observe that certainty factor for inverse decision rules are coverage factors for the original decision rules.

7. What the data are telling us

The above properties of decision tables (algorithms) give a simple method of drawing *conclusions* from the data and giving *explanation* of obtained results.

From the decision algorithm and the certainty factors we can draw the following conclusions.

- No churn is implied with *certainty* by:
  - high number of incoming calls,
  - low number of incoming calls and low number of outgoing calls to other mobile operator.
- Churn is implied with *certainty* by:
  - high number of outgoing calls to other mobile operator,
  - medium number of incoming calls and low number of outgoing calls.
- Clients with medium number of incoming calls and low number of outgoing calls within the same operator are *undecided* (no churn, cer. = 0.48; churn, cer. = 0.52).

From the inverse decision algorithm and the coverage factors we get the following explanations:

- the *most probable* reason for no churn is low general activity of a client,
- the *most probable* reason for churn is medium number of incoming calls and medium number of outgoing calls within the same operator.

8. Summary

In this paper the basic concepts of rough set theory and its application to drawing conclusions from data are discussed. For the sake of illustration an example of churn modeling in telecommunications is presented.

References

[1] J. Grant, "Churn modeling by rough set approach", manuscript, 2001.

[2] S. K. Pal and A. Skowron, Eds., *Rough Fuzzy Hybridization*. Springer, 1999.

[3] Z. Pawlak, *Rough Sets – Theoretical Aspects of Reasoning about Data.* Boston, London, Dordrecht: Kluwer, 1991.

[4] Z. Pawlak, "Decision rules, Bayes' rule and rough sets", in *New Direction in Rough Sets, Data Mining, and Granular-Soft Computing*, N. Zhong, A. Skowron, and S. Ohsuga, Eds. Springer, 1999, pp. 1–9.

[5] Z. Pawlak, "New look Bayes' theorem – the rough set outlook", in *Proc. Int. RSTGC-2001*, Matsue Shimane, Japan, May 2001, pp. 1–8; *Bull. Int. Rough Set Soc.*, vol. 5, no. 1/2, 2001.

[6] L. Polkowski and A. Skowron, Eds., *Rough Sets and Current Trends in Computing*. Lecture Notes in Artificial Intelligence 1424, Springer, 1998.

[7] L. Polkowski and A. Skowron, Eds., *Rough Sets in Knowledge Discovery*. Vol. 1–2, Springer, 1998.

[8] L. Polkowski, S. Tsumoto, and T. Y. Lin, Eds., *Rough Set Methods and Applications – New Developments in Knowledge Discovery in Information Systems*. Springer, 2000, to appear.

[9] N. Zhong, A. Skowron, and S. Ohsuga, Eds., *New Direction in Rough Sets, Data Mining, and Granular-Soft Computing*. Springer, 1999.

More info about rough sets can be found at:

http://www.roughsets.org
http://www.cs.uregina.ca/∼roughset
http://www.infj.ulst.ac.uk /staff/I.Duentsch
http://www-idss.cs.put.poznan.pl/staff/slowinski/
http://alfa/mimuw.edu.pl
http://www.idi.ntnu.no/∼aleks/rosetta/
http://www.infj.ulst.ac.uk/∼cccz23/grobian/grobian.html

Zdzisław Pawlak
Institute of Theoretical and Applied Informatics
Polish Academy of Sciences
Bałtycka st 5
44-000 Gliwice, Poland

Paper

Using reporting and data mining techniques to improve knowledge of subscribers; applications to customer profiling and fraud management

Jean-Louis Amat

Abstract — One of the main issues of operators today is to be able not only to store and manage the huge amount of data generated by the applications and customer contact points, but also to give value to these data. But this implies using tools for storing the data, to manage it, look at it, understand it, exploit it, generate actions such as marketing campaigns. It is therefore obvious that using one tool for each of these functions will lead to a too big and unusable solution. We will discuss here the technical issues involved and show how we turned them into an easy-to-use solution for business users.

Keywords — data mining, customer profiling, OLAP reporting, fraud management, CRM, real-time marketing.

1. Introduction

Knowing the subscribers is one of the major issues of today business. It could seem easy to do because we have to deal with a great amount of data sources recording every transaction, preference or behavior pattern. But these data are precisely too numerous and too complex. So if we want business-oriented people to be able to deal with them, it is necessary to provide not only powerful and complex processing methods, but also easy-to-use, and this is where the actual issue is [1].

The data understanding process follows a complex chain in which we have to master every step. It goes from data acquisition to learning and generating knowledge. Here are the main identified steps [2]:

- Data acquisition depends on the contact points with subscribers, where the actual interaction is, at any level: switches (CDR – call datail records), billing, customer care, points of sales, etc. The usable knowledge should be disseminated in all these data sources and the solutions we are talking about here will have to be able to get the valuable data from any of them.
- Data storage: this is a quite technical issue, but very crucial. As we are dealing sometimes with millions of subscribers, over time periods ranging from a few months to a few years, important data storage means are needed.
- Data management: this is where the system intelligence should start, allowing users to access easily to what they want, the way they want. Marketing databases today are very often structured around a data mart, using a very convenient representation for automatic requesting.
- Data processing: dealing with mathematical functions able to correlate data, to discover patterns, to compute trends or to predict them. We are in the field of data analysis and statistics, but also data visualization (reporting) which is very important in the process of knowledge discovery.
- Data understanding: getting knowledge from data is the part for what we call data mining. It is a complex process using any of the results obtained in the preceding steps, sometimes requiring new processing, in order to understand deeply the data, to get usable and hopefully unknown knowledge.
- Learning: the last part of the process is the most important one and has to remember what was useful in the data/information/knowledge about the subscriber in order to feed future analysis and understanding process. It is obvious today that this part is not present in most of the existing tools but they will have to deal with this in the future.

We are presenting here two solutions we built in order to apply these techniques to fraud management, and to marketing automation (customer analysis and marketing campaigns management).

2. Getting decisions from data

Usable information that will give knowledge about one subscribers is contained in the data we have about his transactions and history. But it is hopeless to try to use every data available; some choices have to be made depending on the final purpose of the application. If we are interested in churn management, we will have to select relevant data to understand the reasons why people are leaving; this is one of the most complicated phenomena to analyze in this field and the solution should be completely specific. But basically, the behavioral data is very important to look at, as well as demographic and financial information about the subscribers. The profiling analysis process is very complex because whatever the data we can put in the analysis,

we can almost always discover patterns. But are these patterns relevant? This is the main concern one has to deal with. For the fraud analysis, it is well known that the calling patterns are often significant in this respect, so a fraud management tool will have to cope with these data, which means analysis of the CDR.
Customer profiling and behavior analysis often rely on the same data sources, present in the telco world. The mostly used data sources are for example the billing system and the customer care.

- Billing: this is where the usage information is stored, as well as some personal information about the subscribers. Depending on the billing implementation, this information can be very accurate, or more general. A good recommendation would be to think of the further profiling application when designing the system.
- Customer care: data is stored also about customer information, interactions with the operator, and mostly short term data history.

There are several ways to use the data and to provide operational results for marketing purpose or fraud analysis, depending on the users. Back office people need solutions to get quickly and easily reports about their favorite subjects, or deeper analysis of the data/information, while front office management is more concerned with managing customer interaction. This leads to separate functions:

- **Reporting and analysis**. This kind of analysis relies on quite simple tools and solutions, and is mainly based on OLAP modeling (on-line analytical processing). The users are decision makers such as general management, marketing manager, etc. We are actually in a decision support process, but where the intelligent part is completely managed by the user. The only intelligence within the system is the way data has been modeled with multidimensional representation. Nevertheless, this kind of solution is very useful to broadcast automatic reports to decision makers. Alcatel CMI (customer management intelligence) solution includes these features as a first analysis layer.
- **Profiling and segmentation**. The output here is not only information (high values, trends) but also models for prediction or classification. The goal of this layer is to provide more that information, i.e. knowledge of the subscribers. The techniques involved are therefore mainly based on artificial intelligence, and we are using decision trees to generate explicit knowledge (rules), neural networks or Bayesian techniques to model the subscribers buying patterns. Usually, the main issue with these techniques is that users need to understand what they are doing in order to get valuable knowledge; but some algorithms can be predefined with standard parameters in order to provide a global result. Alcatel solutions are in this respect among the most easy to use; CMI solution offers “click and play” data mining features fully integrated, while FMS (fraud management system) uses neural networks in a totally transparent way to generate alarms and help decision.
- **Real-time decision making**. This is the original part of Alcatel offer, relying on E.piphany software. The real-time platform is able, during an interaction with a subscriber, to compute a decision profile in seconds and to push an offer through the selected channel (call center, web, chat, etc), without human intervention. In this case the decision (selecting and proposing an offer) can be taken by the system itself (on a web site for example) or by an agent who can select among several propositions (call center). This intelligent layer includes an automatic real-time learning engine able to make profit of any transaction and to update itself the predictive model. The solution is a combination of rule-based system, self-learning analytics (relying on Bayesian techniques) and collaborative filtering.

The Alcatel CMI solution [3] includes the three presented layers with the first and second ones working for back office people and the last one deployed in front office, facing the subscriber (through an agent on a call center, or directly on the web). The fraud management tool FMS contains also reporting and analysis facilities devoted to fraud alarms.

3. Segmentation, classification and profiling

The functions we described rely on a set of techniques that are able to assist the users in their tasks [4]:

- Segmentation is the process to find classes in the data (Fig. 1).

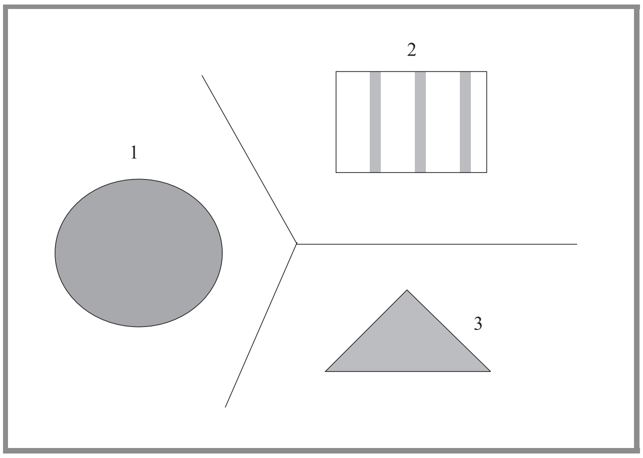


Fig. 1. Classes in the data.

- Classification assumes that a segmentation already exists and tries to attach an element to an existing class or category (Fig. 2).

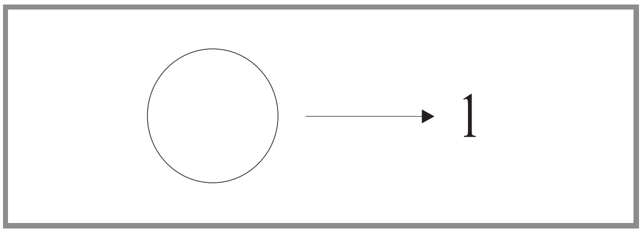


Fig. 2. Classification.

- Profiling consists in describing the elements of a class (Fig. 3).

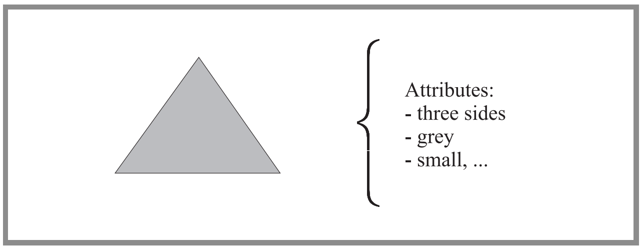


Fig. 3. Profiling.

The main functions to perform when trying to find knowledge about the subscribers often fall into one of these categories. Generally, the process consists in applying a segmentation process to the data and then to classify new data, or to understand the content of the classes. For the example of churn management, we can identify segments in the data representing the different cases of churn. Then we can use profiling techniques to understand each category.

The CMI solution provided by Alcatel contains functions for performing segmentation or profiling, but also predictive models. We give here a short list of the main functions provided [5]:

- **Profiles** is a charting, visualization and description application that enables users to inspect and understand their data visually. Profiles can predict for example how customer segments are varying over time, how product sales vary by region, what the most common profession is for each division, etc.

- **Basic trends** provides time series trending analysis capabilities including straight line growth, constant percentage growth, or moving average trends, including extrapolations of one, two or three periods.

- **High and low clusters** is an anomaly detection method that allows to find the highest- and lowest-performing groups within an attribute or set of attributes according to some measurement. It is very helpful to answer questions such as finding the best and worst customers (the customers who buy much more than expected are probably the best), finding product lines where current year revenue is surprisingly higher than last.

- **Clustering** is a segmentation tool that can be used for customer segmentation, or finding outliers: sometimes the most useful clusters are the smallest ones. These “nuggets” might represent a unique niche or highly profitable (or unprofitable) customers, for example.

- **Influence** is a classification and regression component that is used for two main purposes: first, to find which of the input attributes have the most power to predict the target, and second, to build predictive classification and regression models. Once created, the models can be used to score lists of customers, for example based on the likelihood a customer will respond to a particular campaign.

- **Bayes classifier** allows users to use Naïve or optimal dependency tree (ODT) Bayes classifiers to create classification model, for example helping users to identify characteristics of profitable and unprofitable product lines, select new sales prospects based on the buying patterns of current customers, and perform other types of predictive analyses.

- **Scoring** allows users to use models built in **influence**, **clustering**, **and Bayes classifier**, or use a predefined measure, to rank customers in a list and to target marketing communications where they can be most effective. For example, suppose a user was concerned about high rates of customer attrition. That user could use **influence or Bayes classifier** to build a model to predict which customers were most likely to defect. Then, with the **scoring** application, the user could generate a scored list that ranked customers based on attrition likelihood, and select only the highest risks, in order to address them a specific offer.

4. Fraud management

In a normal day of activity, a telephone company has the potential of creating many hundreds of thousands or even millions of call records. Within this mass of data are calls being made by people who either are targeting the organisation with the aim of defrauding it or people potentially using the service with no intention to pay. And the only way to detect quickly these people is to monitor their usage which is contained in the data calls (CDR). The system Alcatel built for this purpose gets the CDR directly and provide decision support for fraud managers with the help of a rule-based system containing explicit knowledge, call query facilities, usage variation analysis (neural networks application) and also subscriber fingerprinting analysis to help recognizing fraudsters through their calling habits [6] (Fig. 4).

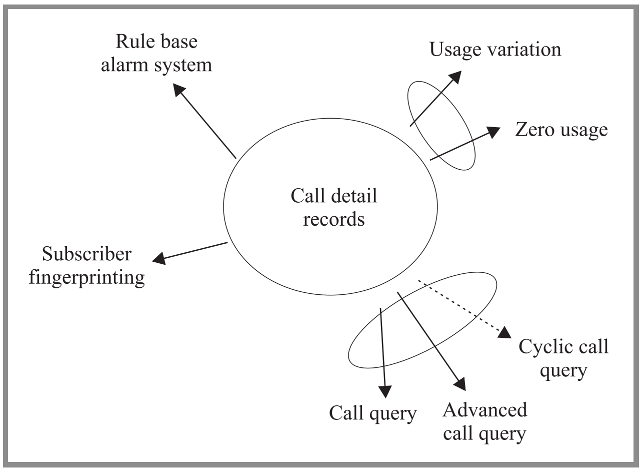


Fig. 4. Various usage of CDR.

When the fraud analysis process commences, the raw call records are sent through a data interface programme for conversion into FMS CDR formats. The system calculates the call duration for each completed call record received, rates the call using an internal tariff table (if CDR are delivered unrated) and merges newly received CDR with existing CDR. When all new CDR have been processed, the system compares the activity per subscriber against triggers established in the system and generates an alarm report each time the set threshold is exceeded by any of the prescribed amounts.

4.1. Call query

The data held within the FMS is very high volume; many million records per day can be easily managed by the system. The call query module allows simple interrogation of that data. By entering one or more search criteria, the user can mine for data. The resulting analysis itself could be many thousands of records or it could, if the search criteria were precise enough, bring back a single record.

4.2. Velocity checking

When two calls are made from two different locations and with a time difference greater than the minimum time required to travel from one to the other location, this optional FMS feature is able to identify and display the relevant CDR.

4.3. Usage variation

Fraud is often detected in its earliest forms by the change in usage of a subscriber. Very little fraud of any substance can be undertaken without usage accelerating on a subscriber's account. The usage variation module of the FMS can assist the user to detect major changes in usage very quickly. The FMS usage variation module is implemented as a feed-forward neural network, fed by input vectors which represent values of parameters assumed by each individual subscriber in the previous 24 hours. The input elements of the vectors are weighted to produce a scalar output value that represents the status of the subscriber for the given day. Over a certain period of time, a reference output value is stored into an alarm reference file, this reference being the maximum output value assumed by the subscriber. The system is thus capable of generating an alarm whenever a subscriber has an output value that exceeds the relevant stored maximum.

4.4. Subscriber fingerprinting

This optional functionality has been introduced in order to address the issue of subscription fraud. In this type of fraud, individuals would subscribe with a telecom operator under false identification. Bills, possibly after a first "quiet" period of normal behaviour, are eventually not paid. After disconnection, they again register using different identification. A possible solution to this problem lies with the assumption that the fraudster should have a specific pattern, like a **signature**, which is defined by some of the numbers called. These numbers could be the majority of his/her calls, or just a few calls, but with a certain degree of **uniqueness**, which might permit to identify the fraudster. **Fingerprints** (identifying patterns), stored in an internal database, made of sequences of called numbers from specified accounts are matched against new account traffic.

5. Customer analysis

Using raw data such as CDR with some explicit knowledge can help to perform powerful fraud management. But for marketing purposes, it is needed to analyze more sophisticated data, running from billing information to clickstream analysis over a web site. In order to get these data and to understand them, the CMI solution is able to store it into a data mart and to provide a fully integrated interface to activate together reporting, data mining, and active marketing campaign management.

For example, suppose a marketing user is concerned about high rates of churn. That user could use **influence** to build a model using historical data to predict which customers were most likely to defect to a competitor. Then, with the scoring application, the user could generate a scored list that ranked customers based on how likely each were to defect. Finally, the user could import that scored list into the campaign manager module to direct a special marketing incentive at the most likely victims of attrition. Within CMI solution, every described operation is performed with the same tool, giving the facilities to switch from OLAP to data mining, then to campaign management, and back to OLAP if needed (Fig. 5).

For performing churn analysis, for example, the basic idea is to detect in a first time the people that has churned in the

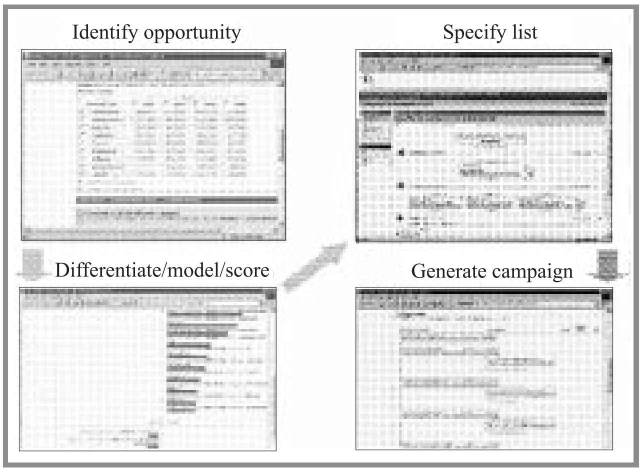


Fig. 5. The process of generating list of clievts for marketing compaign.

past, within a specific time range (to be defined). Regarding this, the system helps to extract a list of "churners" from the customer database, which will be used to feed the learning functions. The data mining tools can then be applied on this population, in order to assess two models.

The first model computed is a **scoring model**. The model will compute from existing/known churners a way to assign a score to each subscriber, which will depend on selected attributes. Once established, this scoring model will be applied on active subscribers and will therefore assess a propensity to churn. Sorting the list of subscribers with this specific measure will give the most probable churners. It will then be possible to activate a specific retention campaign in order to keep these subscribers.

The second model is a **segmentation model**.

In fact, it might be interesting, before acting on the global list of potential churners, to detect the most profitable ones, and to estimate the global revenue regarding the cost of the campaign. The segmentation model will provide a profiling of the churn population, which will have to be interpreted by the analysts (Fig. 6).

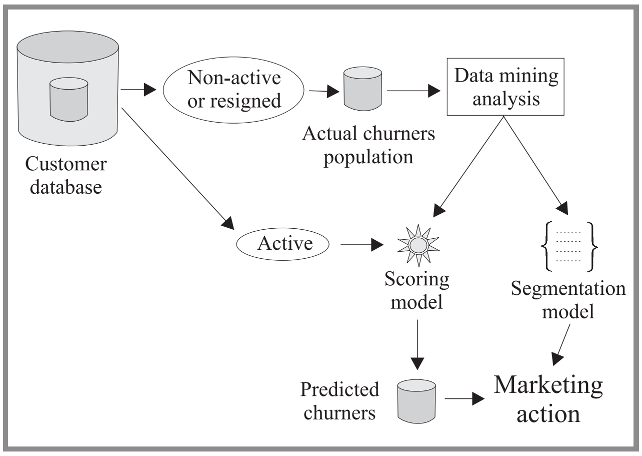


Fig. 6. Segmentation model.

6. Conclusion

Using a combination of the functions described, we are able to propose a fully integrated and operational solution. The decision support part is also integrated with marketing automation and CRM (customer relationship management). This allows to get a unique view of the subscribers, through any interaction channel (Fig. 7).

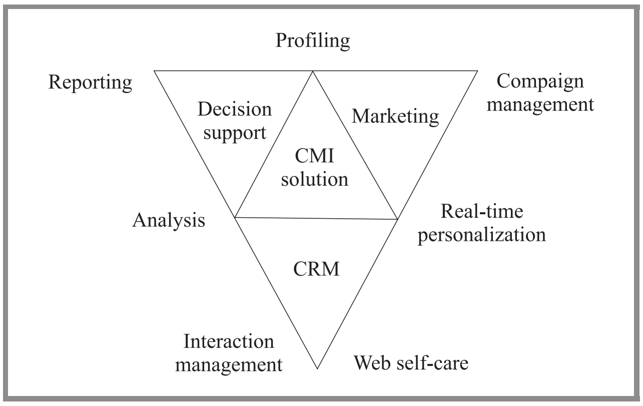


Fig. 7. Integrated solution.

Dealing with complex and heterogeneous data sources is an actual issue in order to perform tasks such as fraud analysis or customer profiling in order to activate marketing campaign.

Alcatel solutions we presented are able to provide already developed interfaces to the standard data sources, and integrated functions for powerful analysis to support decision for fraud detection, customer data management and analysis, and customer relationship management [7].

References

[1] A. Berson, S. Smith, and K. Threaling, *Building Data Mining Applications for CRM*. McGraw Hill, 2000.

[2] J. L. Amat and G. Yahiaoui, "Advanced techniques for information processing: neural networks, fuzzy logic, genetic algorithms", Cepadues Eds., 1995 (in French).

[3] Alcatel, "Customer management intelligence solution", 2000.

[4] J. L. Amat, "Applying customer profiling and segmentation methods", in *IIR Sem. Customer Profiling & Segmentation for Telcos*, London, July 2000.

[5] E.piphany, "E.4 Datamining overview", 1999.

[6] Alcatel, "Alma FMS: fraud management system", 1999, 2000.

[7] J. L. Amat, "Latest CRM and eCRM trends for telcos", in *IIR Sem. eCRM*, London, Feb. 2001.

Jean-Louis Amat
e-mail: Jean-Louis.Amat@alcatel.fr
Network Applications Division
Alacatel
1, rue Ampere
91302 Massy cedex, France

Paper

Analytical properties of a stochastic teletraffic system with MMPP input and an access function

Lino Tralhão, José Craveirinha, and Domingos Cardoso

Abstract — Stochastic modeling of teletraffic systems with restricted availability and correlated input arrival rates is of great interest in GoS (grade of service) analysis and design of certain telecommunication networks. This paper presents some analytical properties of a recursive nature, associated with the infinitesimal generator of the Markov process which describes the state of a teletraffic system with MMPP (Markov modulated Poisson process) input traffic, negative exponentially distributed service times, finite queue and restricted availability defined through a loss function. Also the possible application of the derived properties to a direct method of resolution of the linear system, which gives the stationary probability distribution of the system, will be discussed.

Keywords — stochastic analysis of telecommunication networks, teletraffic theory, GoS analysis of overflow teletraffic systems, queuing systems.

1. Introduction

Problems of performance analysis in telecommunication networks led in the past to the concept of restricted availability systems in which the connection paths may be such that an incoming call may be unsuccessful even when there are still idle circuits in the destination group. In classical studies [1] of teletraffic link systems "loss functions" were used to represent in simple mathematical terms the effects of the restricted availability with respect to the arriving calls for service. This function $\big(w(\upsilon)\big)$ is defined as the conditional probability that a call arriving when there are υ occupied servers, is rejected. In particular this concept was used for calculating the blocking probability of restricted availability overflow systems arising in teletraffic networks with alternative routing. Although these systems typically did not have queuing facilities, modern technologies may provide systems with limited waiting room (say k queuing positions in a buffer). We also may consider teletraffic systems where decisions regarding the acceptance or rejection of a call are of a probabilistic nature and based on the number of calls already in progress (see example in [2]) or waiting for service, mechanism which could be also represented by some specific type of loss function. An example could be the case of "load sharing" [3] schemes of adaptive dynamic routing in multiexchange networks in which calls rejected by a given route are offered to alternative routes according to a set of probabilities which are a function of the states (number of occupations) of the individual groups of channels in the different links of the network.

On the other hand a number of studies [4–7] suggest that the MMPP could be used successfully for modeling certain types of superposition of complex teletraffic flows, including packetized voice and packet data traffic as well as video sources traffic in ATM networks. In particular the MMPP is the exact model for the superposition of independent IPP (interrupted Poisson processes), representing overflow traffics resulting from the overflow of Poisson inputs in loss systems with exponential distribution of the service times (model of great interest in circuit-switched networks with alternative routing).

The m-MMPP point process may be defined as a doubly stochastic Poisson process where the intensity process $\{\lambda(t),\, t \geq 0\}$ is governed by an ergodic Markov process, with m states, i.e.:

$$\lambda(t) := \lambda_{I(t)}\,,$$

where the R.V. (random variable) $I(t)$ indicates the state, at instant t, of an ergodic Markov process. When $I(t) = f$, $f = 1, \ldots, m$, the MMPP is said to be in phase f.

The MMPP is also a particular case of the "Versatile Markovian Point Process" model in [8] and may also be treated as a particular case of the Markovian arrival process model, see [9] and [10].

In a previous work [11], the exact analysis of a loss system with a m-MMPP input, a finite queue of capacity k, N servers with negative exponential service times and a loss function $\omega(\upsilon) := 1-\alpha_\upsilon$, was performed. The extension of this work by considering the exact analysis of a system with finite queuing capacity whose inputs are defined from a number of independent MMPPs each being subject to a particular "access function" is given in [12].

The analysis of such systems, including the characterization of the associated key processes (describing the system state, the overflow traffic and the carried traffic) is expressed in terms of the infinitesimal generator of the Markov process which describes the state of the system, Q. This paper presents some analytical properties of a recursive nature, associated with that infinitesimal generator. The considered loss function of the system may in general depend both on the number of occupations and the phase of the input MMPP. The paper begins by reviewing the basic features of the ergodic Markov process which represents the

$$Q=\left[\begin{array}{lllllll}
A=\underline{\alpha}(0)\Lambda & \alpha(0)\Lambda & & & & & \\
\mu I & A-\underline{\alpha}(1)\Lambda-\mu I & \underline{\alpha}(1)\Lambda & & & & \\
 & & \cdots & & & & \\
 & & N\mu I & A-\underline{\alpha}(N)\Lambda-N\mu I & \underline{\alpha}(N)\Lambda & & \\
 & & & N\mu I & A-\underline{\alpha}(N+1)\Lambda-N\mu I & \underline{\alpha}(N+1)\Lambda & \\
 & & & & & \cdots & \\
 & & & & & N\mu I & A-N\mu I
\end{array}\right] \qquad (2)$$

system state by describing the structure of its infinitesimal generator Q obtained from previous work of the authors. Next, some recursive formulae for the matrix which contains the basis of the vector space of the solutions associated with a submatrix of a matrix of the type of Q, are derived by exploring the diagonal block structure of this type of matrices. These properties are then applied to Q having in mind its specific block structure. Also the possible application of the derived properties to a direct method of resolution of the linear system which gives the stationary probability distribution of the system, will be considered. Some numerical examples of application of such a direct method will be presented in order to illustrate its potential advantages and limitations.

2. Characterization of the system

Let us consider the stochastic service system represented in Fig. 1, with m-MMPP input, finite queue k, N servers, negative exponentially distributed service times (with mean μ^{-1}) and a loss function $\omega(\upsilon,f)=1-\alpha(\upsilon,f)$, $\upsilon=0,\dots,N+k$, $f=1,\dots,m$.

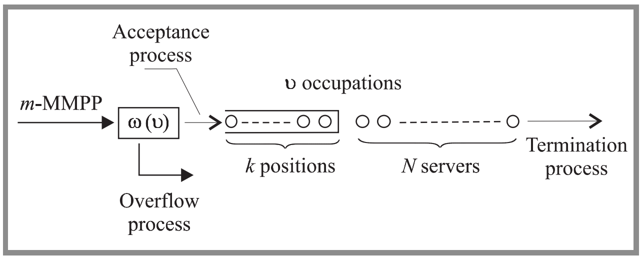


Fig. 1. Stochastic service system.

Note that the input m-MMPP may represent itself the superposition of a number of independent m_r-MMPPs, and the access function $\alpha(\upsilon,f)$ enables to represent the conditional probability of an arrival being accepted when the system is in state (υ,f), where υ is the number of occupations and f is the current phase of the input process (this general case was analysed in [12]), assuming that each m_r-MMPP has a particular access function $\alpha_r(\upsilon)$. The details of the analysis of the system, namely the characterization of the overflow process, the acceptance process and the termination process are given in [11] and [12].

The stochastic process $\{X_t,t\geq 0\}$ which describes the system state at instant t, has the state space:

$$I=\{i=(\upsilon,f),\ \ \upsilon=0,\dots,N+k,\ \ f=1,\dots,m\} \qquad (1)$$

and is an ergodic Markov process. X_t is characterized by the infinitesimal generator [12] – see Eq. (2), shown at the top of this page, where:

$$\underline{\alpha}(\upsilon)=\operatorname{diag}\bigl(\alpha(\upsilon,1),\dots,\alpha(\upsilon,m)\bigr)\,,$$
$$\Lambda=\operatorname{diag}(\lambda_1,\dots,\lambda_m)\,,$$

λ_f is the intensity of the input MMPP in phase f and A is the infinitesimal generator of the ergodic Markov process governing the intensity process of the input MMPP.

An essential element of the system analysis or of any system with similar infinitesimal generator is the stationary measure π of Q (stationary probability distribution):

$$\pi=\bigl[\pi_i\bigr],\quad i\in I\,, \qquad (3)$$

such that:

$$\left\{\begin{array}{l}\pi Q=0\\ \pi e=1\end{array}\right.\,,$$

where e is the column matrix $e=[1,\dots,1]^T$.

Note that the most relevant GoS parameters of this type of system, namely the call congestion and the waiting probability, may be expressed in terms of π (see [11] and [12]). In this paper a recursive formula for π will be derived which beyond its analytical value may also be used for a direct resolution of the linear system (3).

3. Analytical properties

3.1. Preliminary analysis and results

Let us consider the linear system (3), where Q is the square matrix composed of $S+1$ rows (and columns) of square blocks of order m:

$$Q:=\left[\begin{array}{lllll}
A_0 & C_0 & & & \\
M_1 & A_1 & C_1 & & \\
 & & \cdots & & \\
 & & M_i & A_i & C_i \\
 & & & \cdots & \\
 & & & M_S & A_S
\end{array}\right]. \qquad (4)$$

It is assumed that Q is the infinitesimal generator of an ergodic jump Markov process (of stationary measure π), and the matrices $M_1, \dots, M_S$ are regular.

Let us now consider the submatrix Q' with $(S+1)$ block rows and S block columns, obtained from Q by eliminating the block column S:

$$Q' := \left\lfloor \begin{matrix} A_0 & C_0 & & & & \\ M_1 & A_1 & C_1 & & & \\ & & \dots & & & \\ & & M_i & A_i & C_i & \\ & & & & \dots & \\ & & & & M_{S-1} & A_{S-1} \\ & & & & & M_S \end{matrix} \right\rfloor . \quad (5)$$

Since the square submatrix Q'':

$$Q'' := \left\lfloor \begin{matrix} M_1 & A_1 & C_1 & & \\ & & \dots & & \\ & & M_{S-2} & A_{S-1} & C_{S-2} \\ & & & M_{S-1} & A_{S-1} \\ & & & & M_S \end{matrix} \right\rfloor \quad (6)$$

of order $S.m$ is regular (because all its diagonal blocks $M_1, \dots, M_S$, are regular), we conclude that Q' has also rank $S.m$ (Q' has $S.m$ columns). Therefore the set s of solutions of the homogeneous system:

$$uQ' = 0 \quad (7)$$

constitutes a vector space of dimension $(S+1)m - Sm = m$. Let $T := \{x_0, \dots, x_{m-1}\}$ be a basis of s, and consider the rectangular matrix:

$$X := \left\lfloor \begin{matrix} x_0 \\ \dots \\ x_{m-1} \end{matrix} \right\rfloor = \left[X_0 \,|X_1|\, \dots \,|X_S\right], \quad (8)$$

where every square block X_i has order m.

From Eqs. (7) and (8):

$$XQ' = 0 \,. \quad (9)$$

Obviously $\pi \in s(\pi Q = 0 \Rightarrow \pi Q' = 0)$. Then we have:

$$\pi = \gamma X \,, \quad (10)$$

where $\gamma := [\gamma_0, \dots, \gamma_{m-1}]$ is a row matrix of order m.

Representing by I_k the identity matrix of order k, putting:

$$X_0 = I_m \quad (11a)$$

(which guaranties T as a basis of s) and multiplying the first block column of Q' by X we obtain:

$$X_0A_0 + X_1M_1 = 0 \Leftrightarrow X_1 = -X_0A_0M_1^{-1} = -A_0M_1^{-1}. \quad (11b)$$

Using now the second block column:

$$X_0C_0 + X_1A_1 + X_2M_2 = 0 \Leftrightarrow X_2 = -(X_0C_0 + X_1A_1)M_2^{-1}. \quad (11c)$$

For the $(i-1)$th block column:

$$X_i = -(X_{i-2}C_{i-2} + X_{i-1}A_{i-1})M_i^{-1}, \quad i = 2, \dots, S. \quad (11d)$$

Therefore Eqs. (11) allow us to obtain X recursively.

Let us now designate by Q_i, $i = 0, \dots, S$, the ith block column of matrix Q. From (9) we have $XQ_i = 0$, $i \in \{0, \dots, S-1\}$. However the ergodicity of the Markov jump process referred to above implies, as is well known, that the set of solutions of Eq. (3) constitutes a vector space of dimension 1. Then we have, for $m > 1$ (the case $m = 1$ is trivial because π becomes directly determined from (11)):

$$XQ_S \neq 0 \,. \quad (12)$$

Therefore (12) in conjunction with Eqs. (10) and (3) allows us to say that γ can be obtained by resolving the m-order system:

$$\gamma(XQ_S) = 0 \,. \quad (13)$$

Explicitly:

$$\gamma(X_{S-1}C_{S-1} + X_SA_S) = 0 \,. \quad (14)$$

This system is obviously singular but its vector space of solutions has dimension 1. By introducing γ in (10), π becomes known in terms of X after the normalization $\pi e = 1$.

Remark. The present analysis is also applicable to systems with Q having the same general properties, but with the form:

$$Q = \left\lfloor \begin{matrix} A_{00} & A_{01} & A_{02} & \dots & A_{0S} \\ M_1 & A_{11} & A_{12} & \dots & A_{1S} \\ & M_2 & A_{22} & \dots & A_{2S} \\ & & & \dots & \\ & & & M_S & A_{SS} \end{matrix} \right\rfloor . \quad (15)$$

In this case, (11) becomes:

$$X_i = -\left(\sum_{j=0}^{i-1} X_jA_{j,i-1}\right) M_i^{-1}, \; i = 1, 2, \dots, S. \quad (16)$$

3.2. Application to the system

In the case of the matrix Q of the system shown in Fig. 1 we have $S = N + k$ and:

$$\begin{aligned} M_i &= f(i)\mu I_m, & i &= 0, \dots, S \\ C_i &= \underline{\alpha}(i)\Lambda, & i &= 0, \dots, S-1 \\ C_S &= 0 & & \\ A_i &= A - M_i - C_i, & i &= 0, \dots, S, \end{aligned} \quad (17)$$

where $\alpha(i)$ and Λ are diagonal of order m, I_m is the identity matrix of order m, A is singular of order m and:

$$f(i) := \begin{cases} i & \text{if} \quad i < N \\ N & \text{if} \quad i \geq N \end{cases} .$$

Proposition. For the case of matrix Q (2) and making $X_0 = I_m$, X is given recursively by:

$$X_i = \left\lfloor X_{i-1}C_{i-1} - \left(\sum_{j=0}^{i-1} X_j\right) A \right\rfloor M_i^{-1},$$
$$i = 1, \ldots, N+k. \tag{18}$$

Proof (by induction):

1. From Eqs. (11b) and (16) (for $i = 1$):
$$X_1 = -A_0 M_1^{-1} = (C_0 - A) M_1^{-1}$$
which satisfies Eq. (18), taking into consideration that $X_0 = I_m$.
2. From Eq. (11d):
$$X_{i+1} = -\left(X_{i-1}C_{i-1} + X_i A_i\right) M_{i+1}^{-1}. \tag{19}$$
Substituting (17) in (19):
$$X_{i+1} = \left(-X_{i-1}C_{i-1} - X_i A + X_i C_i + X_i M_i\right) M_{i+1}^{-1}.$$
Introducing (18):
$$X_{i+1} = \left\lfloor -X_{i-1}C_{i-1} - X_i A + X_i C_i + X_{i-1}C_{i-1} + \right.$$
$$\left. - \left(\sum_{j=0}^{i-1} X_j\right) A \right\rfloor M_{i+1}^{-1} = \left[X_i C_i - \left(\sum_{0}^{i} X_j\right) A\right] M_{i+1}^{-1}.$$

□

From Eq. (18), with $i = S$ we have:

$$X_S M_S = -\left(\sum_{j=0}^{S-1} X_j\right) A + X_{S-1} C_{S-1}$$

or:

$$\left(\sum_{j=0}^{S-1} X_j\right) A = X_{S-1} C_{S-1} - X_S M_S \,.$$

By adding $X_S A$ to both sides we obtain:

$$\left(\sum_{j=0}^{S} X_j\right) A = X_{S-1} C_{S-1} - X_S M_S + X_S A$$

or, taking into consideration that $A_S = A - M_S$:

$$\left(\sum_{j=0}^{S} X_j\right) A = X_{S-1} C_{S-1} + X_S A_S$$

this implies, taking (15) into consideration, that the m-order system:

$$\gamma \left[\left(\sum_{j=0}^{S} X_j\right) A\right] = 0 \tag{20}$$

may be used for obtaining γ.

The result (20) can also be derived from stochastic considerations, noting that $u := \sum_{0}^{N+k} \pi_\upsilon$ is the stationary probability measure of the underlying Markov jump process of the input MMPP.

3.3. The case $m = 2$

In this case, formulae (18) and (20) can be simplified. In fact since A is the infinitesimal generator of a Markov jump process, the sum of the elements of each row is zero. In other words, the first and the second columns of A have symmetrical elements. Let

$$A_0 := \left\lfloor \begin{matrix} a_{0,0} \\ a_{0,1} \end{matrix} \right\rfloor, \; A_1 := \left\lfloor \begin{matrix} a_{0,1} \\ a_{1,1} \end{matrix} \right\rfloor,$$
$${}_iR := \sum_{j=0}^{i} X_j, \; i = 0, \ldots, N+k \tag{21}$$

since $A_0 = -A_1$, this implies ${}_iRA_0 = -{}_iRA_1$. So, this kind of symmetry of matrix A is transmitted to the matrices ${}_iRA$ and these matrix products are simplified.

The space of solutions of the singular system (20) has dimension 1. So we may arbitrate $\gamma_0 = 1$ and put:

$$R := {}_{N+k}R := \left\lfloor \begin{matrix} r_{0,0} & r_{0,1} \\ r_{1,0} & r_{1,1} \end{matrix} \right\rfloor,$$
$$RA = \begin{bmatrix} r_{0,0}a_{0,0} + r_{0,1}a_{1,0} & -\left(r_{0,0}a_{0,0} + r_{0,1}a_{1,0}\right) \\ r_{1,0}a_{0,0} + r_{1,1}a_{1,0} & -\left(r_{1,0}a_{0,0} + r_{1,1}a_{1,0}\right) \end{bmatrix} \tag{22}$$

then:

$$\gamma(RA) = [1, \gamma_1] \times$$
$$\times \left\lfloor \begin{matrix} r_{0,0}a_{0,0} + r_{0,1}a_{1,0} & -\left(r_{0,0}a_{0,0} + r_{0,1}a_{1,0}\right) \\ r_{1,0}a_{0,0} + r_{1,1}a_{1,0} & -\left(r_{1,0}a_{0,0} + r_{1,1}a_{1,0}\right) \end{matrix} \right\rfloor = 0$$

and

$$\gamma_1 = -\frac{r_{0,0}a_{0,0} + r_{0,1}a_{1,0}}{r_{1,0}a_{0,0} + r_{1,1}a_{1,0}} = -\frac{r_{0,0} + \alpha r_{0,1}}{r_{1,0} + \alpha r_{1,1}},$$
$$\alpha := \frac{a_{1,0}}{a_{0,0}}. \tag{23}$$

4. Calculation of the probability distribution

An obvious application of the recursive formula (18) is the resolution of the linear system (3).

In [13], an iterative method for solving a system which is a particular case of the one under consideration (with full availability which corresponds to $\alpha(\upsilon, f) = 1$, for all (υ, f), was presented. This method results from the application of the general procedure for constructing iterative methods (see [14], p. 532):

$$\pi' Q = 0 \Leftrightarrow \pi' R = -\pi'(Q - R) \Leftrightarrow$$
$$\Leftrightarrow \pi' = \pi'(Q - R)(-R)^{-1} \tag{24}$$

and

$$\pi'^{(n)} = \pi'^{(n-1)} (Q - R) (-R)^{-1}, \tag{25}$$

where $\pi'^{(n)}$ is the value of π' after the nth iteration. This scheme converges if the spectral radius of $(I-QR)^{-1}$ is less than 1 ([14], theor. 8.2.1).
In [13] it is considered:

$$R := I_{N+k+1} \otimes (A - \Lambda - N\mu I_m), \tag{26}$$

where $\otimes$ represents Kronecker product.

Putting

$$M = -(A - \Lambda - N\mu I_m) \tag{27}$$

then

$$(-R)^{-1} = I_{N+k+1} \otimes M^{-1}. \tag{28}$$

Introducing (27) and (28) in (25) the following iterative method is now obtained for the system (3), with Q given by (2):

$$\begin{cases} \underline{\pi}_0'^{(n)} &= \Big\{ \underline{\pi}_0'^{(n-1)} \Big[\big(I - \underline{\alpha}(0)\big)\Lambda + N\mu I \Big] + \\ & \quad + \underline{\pi}_1'^{(n-1)} \mu I \Big\} M^{-1} \\ \dots \\ \underline{\pi}_i'^{(n)} &= \Big\{ \underline{\pi}_{i-1}'^{(n-1)} \underline{\alpha}(i-1)\Lambda + \underline{\pi}_i'^{(n-1)} \Big[\big(I - \underline{\alpha}(i)\big)\Lambda + \\ & \quad + (N-i)\mu I \Big] + \underline{\pi}_{i+1}'^{(n-1)} (i+1)\mu I \Big\} M^{-1} \\ \dots \\ \underline{\pi}_N'^{(n)} &= \Big\{ \underline{\pi}_{N-1}'^{(n-1)} \underline{\alpha}(N-1)\Lambda + \underline{\pi}_N'^{(n-1)} \Big[\big(I - \underline{\alpha}(N)\big)\Lambda \Big] + \\ & \quad + \underline{\pi}_{N+1}'^{(n-1)} N\mu I \Big\} M^{-1} \\ \dots \\ \underline{\pi}_{N+j}'^{(n)} &= \Big\{ \underline{\pi}_{N+j-1}'^{(n-1)} \underline{\alpha}(N+j-1)\Lambda + \underline{\pi}_{N+j}'^{(n-1)} \times \\ & \quad \times \Big[\big(I - \underline{\alpha}(N+j)\big)\Lambda \Big] + \underline{\pi}_{N+j+1}'^{(n-1)} N\mu I \Big\} M^{-1} \\ \dots \\ \underline{\pi}_{N+k}'^{(n)} &= \Big\{ \underline{\pi}_{N+k-1}'^{(n-1)} \underline{\alpha}(N+k-1)\Lambda + \underline{\pi}_{N+k}'^{(n-1)} \Lambda \Big\} M^{-1}. \end{cases} \tag{29}$$

As initial value, analogously to Meier [13], we may put:

$$\pi'^{(0)} = \big[(N+k+1).m\big]^{-1} e^T. \tag{30}$$

As an alternative one might apply the recursive scheme (18) for constructing a direct method of resolution of the system:

1. $X_0 = I_m$.
2. For $i = 1, \dots, S$, apply the recursion (18) in X_i.
3. Solve:

   $$\gamma\big(X_{S-1}\underline{\alpha}(S-1)\Lambda + X_{S-1}(A - M_S)\big) = 0$$

   with respect to γ, by any suitable method.
4. Compute $\pi' = \gamma X$ and finally $\pi = \dfrac{\pi'}{\pi' e}$.

This method has the disadvantage of any direct method: error propagation. However it has the advantage of its simplicity and efficiency in terms of implementation, which makes it attractive for systems with small dimension. This method may also be used to obtain a first approximate solution, which may then be improved through an iterative scheme such as (29). Note, on the other hand, that the method takes advantage of the particular block structure of Q.
For an interesting overview of numerical techniques for the resolution of sparse linear systems namely related to Markov processes analysis, see [15].

5. Computational experiments

In Table 1 some computational results are presented, obtained under the following conditions:

$$\mu = 1,\ k = 0,\ N = 160,\ m = 2,$$
$$\alpha(\upsilon) = \operatorname{diag}(1 - c^{N-\upsilon}, \dots, 1 - c^{N-\upsilon}),$$
$$\upsilon = 0, \dots, N-1,\ c = \frac{\lambda}{\mu N}$$

(where λ is the mean intensity of the input m-MMPP, and the choice of $\underline{\alpha}(\upsilon)$ corresponds to the classical "geometric group" approximation by Smith [17]),

$$A = \begin{bmatrix} -a_0 & a_0 \\ a_1 & -a_1 \end{bmatrix},\ \Lambda = \begin{bmatrix} l_0 & \\ & l_1 \end{bmatrix}.$$

Each row corresponds to a calculation of π by three different methods: using recursive formula only (column "recurs"), recursive formula refined by the iterative method (columns "refined" and "nitd") and iterative method only (columns "iterat" and "nitm"). Iterative schemes are stopped when $\max\Big\{\big|\pi_i^{(n)} - \pi_i^{(n-1)}\big|,\ i \in I\Big\} \le 10^{-6}$ (columns "nitd" and "nitm" present the number of iterations in the respective case). After calculation of π, the vector err $= \pi Q$ is evaluated; columns "recurs", "refined", "iterat" present the maximum absolute values of this vector in the three cases:

$$\left.\begin{array}{l} \text{recurs} \\ \text{refined} \\ \text{iterat} \end{array}\right\} = \max\Big\{\big|\text{err}_i\big|,\ i \in I\Big\} = \varepsilon_{\max}.$$

It can be seen that the recursion is sensitive to the "jitter" [16] of the input MMPP. In fact greater values of a_0 and a_1 (which imply increased "jitter") increases the recursion fragility, leading to unacceptable $\varepsilon_{\max}$ unless the refinement through the iterative procedure is applied. Another point to take into consideration concerns the relative values of l_0 and l_1; when l_0 approximates l_1, the input MMPP approximates the Poisson process and recursion efficiency increases. To illustrate this behavior some examples are shown where the input MMPP degenerates into a Poisson process ($l_0 = l_1$); in such examples $\varepsilon_{\max} = 0$. In the great majority of cases the recursion followed by

Table 1
Computational results

N	l_0	l_1	a_0	a_1	λ	recurs	refined	nitd	iterat	nitm
160	80	0	0.1	0.1	40.0	4.8×10^{-2}	2.4×10^{-4}	62	1.6×10^{-4}	1864
160	80	0	0.1	1	72.7	1.1×10^{-3}	1.6×10^{-4}	3	2.8×10^{-4}	1380
160	80	0	0.1	10	79.2	0	2.4×10^{-12}	1	2.8×10^{-4}	997
160	80	0	1	0.01	0.8	3.4×10^{-1}	1.6×10^{-4}	2069	1.3×10^{-4}	2832
160	80	0	1	0.1	7.3	1.0×10^{0}	1.6×10^{-4}	1743	1.3×10^{-4}	2557
160	80	0	1	1	40.0	5.0×10^{-1}	1.6×10^{-4}	997	1.6×10^{-4}	1367
160	80	0	1	10	72.7	0	3.9×10^{-8}	1	2.8×10^{-4}	929
160	80	0	10	0.1	0.8	3.1×10^{1}	1.6×10^{-4}	2166	1.3×10^{-4}	2116
160	80	0	10	1	7.3	2.4×10^{1}	1.3×10^{-4}	1931	1.3×10^{-4}	1823
160	80	0	10	10	40.0	3.0×10^{1}	2.2×10^{-4}	1304	2.3×10^{-4}	1082
160	80	48	0.1	0.1	64.0	0	8.9×10^{-16}	1	2.1×10^{-4}	1148
160	80	48	0.1	1	77.1	0	2.2×10^{-15}	1	2.5×10^{-4}	1352
160	80	48	0.1	10	79.7	0	1.8×10^{-15}	1	2.5×10^{-4}	1006
160	80	48	1	0.1	50.9	0	8.9×10^{-16}	1	1.9×10^{-4}	1629
160	80	48	1	1	64.0	0	6.7×10^{-16}	1	2.1×10^{-4}	992
160	80	48	1	10	77.1	0	1.8×10^{-15}	1	2.5×10^{-4}	980
160	80	48	10	0.1	48.3	0	3.1×10^{-9}	1	1.9×10^{-4}	1415
160	80	48	10	1	50.9	0	3.8×10^{-8}	1	1.9×10^{-4}	1364
160	80	48	10	10	64.0	0	1.9×10^{-7}	1	2.4×10^{-4}	1071
160	80	80	0.1	0.1	80.0	0	8.9×10^{-16}	1	2.4×10^{-4}	931
160	80	80	0.1	1	80.0	0	1.3×10^{-15}	1	2.4×10^{-4}	1098
160	80	80	0.1	10	80.0	0	8.9×10^{-16}	1	2.4×10^{-4}	1013
160	80	80	1	0.1	80.0	0	2.2×10^{-15}	1	2.4×10^{-4}	1098
160	80	80	1	1	80.0	0	6.7×10^{-16}	1	2.4×10^{-4}	931
160	80	80	1	10	80.0	0	8.9×10^{-16}	1	2.4×10^{-4}	1003
160	80	80	10	0.1	80.0	0	8.9×10^{-16}	1	2.4×10^{-4}	1013
160	80	80	10	1	80.0	0	1.3×10^{-15}	1	2.4×10^{-4}	1003
160	80	80	10	10	80.0	0	4.8×10^{-13}	1	2.4×10^{-4}	931
160	160	0	0.1	0.1	80.0	1.1×10^{2}	1.2×10^{-4}	7175	1.6×10^{-4}	1680
160	160	0	0.1	1	145.5	3.0×10^{0}	3.0×10^{-4}	1091	4.0×10^{-4}	1382
160	160	0	0.1	10	158.4	1.4×10^{-2}	2.9×10^{-4}	44	4.2×10^{-4}	996
160	160	0	1	0.1	14.5	3.0×10^{1}	1.7×10^{-4}	2583	1.2×10^{-4}	3018
160	160	0	1	1	80.0	1.2×10^{2}	1.2×10^{-4}	1542	1.7×10^{-4}	963
160	160	0	1	10	145.5	2.0×10^{1}	5.5×10^{-4}	490	4.0×10^{-4}	1176
160	160	0	10	0.1	1.6	1.4×10^{2}	8.9×10^{-5}	2294	1.1×10^{-4}	2160
160	160	0	10	1	14.5	7.4×10^{1}	1.7×10^{-4}	1809	1.2×10^{-4}	1779
160	160	0	10	10	80.0	2.3×10^{1}	2.4×10^{-4}	1180	3.2×10^{-4}	631
160	160	96	0.1	0.1	128.0	0	1.9×10^{-9}	1	2.5×10^{-4}	1220
160	160	96	0.1	1	154.2	0	1.2×10^{-14}	1	3.5×10^{-4}	1383
160	160	96	0.1	10	159.4	0	1.3×10^{-15}	1	3.5×10^{-4}	1429
160	160	96	1	0.1	101.8	0	5.4×10^{-8}	1	2.3×10^{-4}	1771
160	160	96	1	1	128.0	0	1.0×10^{-8}	1	3.2×10^{-4}	1266
160	160	96	1	10	154.2	0	3.6×10^{-15}	1	3.4×10^{-4}	1082
160	160	96	10	0.1	96.6	8.3×10^{-1}	2.5×10^{-4}	692	2.3×10^{-4}	1277
160	160	96	10	1	101.8	1.5×10^{0}	2.5×10^{-4}	826	2.3×10^{-4}	1328
160	160	96	10	10	128.0	1.1×10^{-4}	5.6×10^{-5}	1	3.2×10^{-4}	1346
160	160	128	0.1	0.01	130.9	0	2.2×10^{-15}	1	3.1×10^{-4}	6168
160	160	128	0.1	0.1	144.0	0	1.3×10^{-15}	1	2.9×10^{-4}	1277
160	160	128	0.1	1	157.1	0	2.7×10^{-15}	1	3.3×10^{-4}	1380
160	160	128	0.1	10	159.7	0	1.3×10^{-15}	1	3.3×10^{-4}	1860
160	160	128	1	0.1	130.9	0	1.8×10^{-15}	1	2.7×10^{-4}	1449
160	160	128	1	1	144.0	0	1.8×10^{-15}	1	3.2×10^{-4}	1180
160	160	128	1	10	157.1	0	2.7×10^{-15}	1	3.3×10^{-4}	1083
160	160	128	10	0.1	128.3	0	3.6×10^{-15}	1	2.7×10^{-4}	1615
160	160	128	10	1	130.9	0	9.2×10^{-15}	1	2.7×10^{-4}	1568
160	160	128	10	10	144.0	0	3.1×10^{-15}	1	3.2×10^{-4}	1195

the iterative procedure performs more efficiently then the "pure" iterative procedure. The "refined" recursion tends to be less efficient then the "pure" iterative method when intensity l_1 is close to 0, corresponding to the MMPP "degenerating" into a IPP and when the "jitter" has a significant increase, leading to a direct solution with great error. Many other computational experiments have confirmed these general trends.

6. Conclusions

Analytical properties of a recursive nature, associated with the infinitesimal generator of jump Markov processes describing certain teletraffic systems having a peculiar diagonal block structure, have been derived. These properties were applied to the infinitesimal generator of a system with MMPP input, negative exponentially distributed service times, finite queue and restricted availability defined through a loss function. The resulting recursive formulae may be applied as a direct scheme for the resolution of the linear system, which gives the stationary probability distribution of the system, in terms of which the main GoS parameters may be expressed. Numerical examples with systems of small dimension, suggest that the method error depends critically on the "jitteriness" of the input MMPP and the arrival intensities. Therefore it is recommended that the derived recursion be used to obtain a first approximate solution to be improved through an iterative scheme. Comparison of this "refined" recursive scheme with the "pure" iterative model in [13] indicate that the former performs more efficiently in most cases when the arrival intensities are all relatively far from zero and when the "jitterness" factor of the input MMPP is limited.

Finally note that the obtained recursive formulae are valid for any infinitesimal generator with the considered block structure. Possible application of the recursion to other Markovian stochastic systems with the same type of block structure might be envisaged as future work.

References

[1] B. Wallstrom, "Congestion studies in telephone systems with overflow facilities", *Ericsson Techn.*, no. 3, pp. 190–351, 1966.

[2] A. Arvindson, "On the performance of a circuit switched link with priorities", *IEEE J. Select. Areas Commun.*, vol. 9, no. 2, 1991.

[3] G. Bel, P. Chemouil, J. M. Garcia, F. Le Gall, and J. Bernoussou, "Adaptative traffic routing in telephone networks", *Large Scale Syst. J.*, vol. 8, no. 3, pp. 267–282, 1985.

[4] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetised voice and data traffic and related statistical multiplexer performance", *IEEE Select. Areas Commun.*, SAC-4, pp. 856–868, 1986.

[5] J. M. Holtzman, "Characteristics of superpositions of traffic streams", in *Brussels Spec. Sem. ISDN Traffic Issues*, Brussels, 1986.

[6] M. F. Neuts, "Modelling data traffic streams", in *Proc. 13th Int. Teletraffic Congr.*, Pub. Elsev. Sci., Eds. Jensen and Iversen, 1991.

[7] S. Wang and J. Silvester, "An approximate model for performance evaluation of real time multimedia communication systems", *Perform. Eval.*, no. 22, pp. 239–256, 1995.

[8] M. F. Neuts, "A versatile Markovian point process", *J. Appl. Prob.*, no. 16, pp. 764–779, 1979.

[9] M. F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. New York: Marcel Dekker, 1989.

[10] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process", *Stoch. Models*, vol. 7, no. 1, pp. 41–42, 1991.

[11] L. Tralhão, J. Craveirinha, and J. Paixão, "A stochastic system with MMPP input and an access function", *Appl. Stoch. Models Data Anal.*, vol. 9, pp. 279–299, 1993.

[12] L. Tralhão, J. Craveirinha, and J. Paixão, *A Study on a Stochastic System with Multiple MMPP Inputs Subject to Access Functions*. New Progress in Probability and Statistics. VSP-Inter. Publ., 1994, pp. 415–428.

[13] K. S. Meier, "The analysis of a queue arising in overflow models", *IEEE Trans. Commun.*, vol. 37, no. 4, pp. 367–372, 1989.

[14] J. Stoer and R. Burlisch, *Introduction to Numerical Analysis*. Springer Verlag, 1980.

[15] V. A. Barker, "Numerical solution of sparse singular systems of equations arising from ergodic Markov chains", *Stoch. Models*, vol. 5, no. 3, pp. 335–381, 1989.

[16] K. S. Meier, "A statistical procedure for fitting Markov-modulated Poisson processes". Ph.D. dissertation, Univ. Delaware, 1984.

[17] N. M. H. Smith, "More accurate calculations of overflow traffic from crossbar group selectors", *Telecommun. J. Aust.*, vol. 13, no. 6, pp. 472–475, 1963.

Lino Tralhão
INESC, Rua Antero de Quental, 199
3000 Coimbra, Portugal

Department of Mathematics FCTUC
University of Coimbra
Largo D. Dinis, Coimbra, Portugal

José Craveirinha
INESC, Rua Antero de Quental, 199
3000 Coimbra, Portugal

Department of Electrical Engineering Science FCTUC
University of Coimbra - Pólo II
3030 Coimbra, Portugal

Domingos Cardoso
Department of Mathematics
University of Aveiro
3800 Aveiro, Portugal

Paper

Effectiveness of active forgetting in machine learning applied to financial problems

Hirotaka Nakayama and Kengo Yoshii

Abstract — One of main features in financial investment problems is that the situation changes very often over time. Under this circumstance, in particular, it has been observed that additional learning plays an effective role. However, since the rule for classification becomes more and more complex with only additional learning, some appropriate forgetting is also necessary. It seems natural that many data are forgotten as the time elapses. On the other hand, it is expected more effective to forget unnecessary data actively. In this paper, several methods for active forgetting are suggested. The effectiveness of active forgetting is shown by examples in stock portfolio problems.

Keywords — pattern classification, potential method, additional learning, forgetting.

1. Introduction

In many practical problems, e.g., financial investment problems, the situation changes very often over time. In machine learning, therefore, decision rules are needed to adapt for such changeable situations. To this end, additional learning should be made on the basis of new data. One of the authors and his collaborators have reported the effectiveness of additional learning in several machine learning techniques: mathematical programming approach [1], potential method [2] and RBF networks [3–5].

On the other hand, since the rule for classification becomes more and more complex with only additional learning, some appropriate forgetting is also necessary. Although several trials of forgetting in machine learning have been also suggested, they are concernced in such a way that the degree of importance of data decreases over time [3–5]. We call the way of forgetting based only on the time elapse "passive forgetting". However, it seems more effective to forget data which give bad influences to the current judgment. We call this way of forgetting "obstacle data" actively "active forgetting". In this paper, the effectiveness of active forgetting will be proved through some examples in stock portfolio problems.

2. Potential method

To begin with, the potential method suggested by one of the authors *et al.* [2] is reviewed briefly. The idea of potential method is originated from the static electric theory. Another similar method is the restricted Coulomb energy (RCE) classifier by Cooper [6] and Reilly *et al.* [7]. RCE tries to increase the ability of classification by adjusting the radia of hyperspheres which approximate the region of influence of data.

Unlike RCE, however, the potential method adjusts "charge" associated with each data in order to increase the ability of generalization. Each hidden unit corresponds to each teacher's pattern $x_j (j = 1, \cdots, N)$, which has some amount of charge c_j in which the sign depends on which category it belongs. Letting $D(x, x_j)$ denote a distance between x and x_j, the output unit is connected to

$$z(x) = \mathrm{sgn} P(x),$$

where

$$P(x) = \sum_{j=1}^{N} \frac{c_j}{D(x, x_j)}.$$

Here, P is the well known potential function, which sign decided on which category a given test data belongs to.

Note that the potential method can classify each teacher's data x_j $(j = 1, \cdots, N)$ correctly without doing anything, because $P(x_j) = +\infty$ for $c_j > 0$ and $P(x_j) = -\infty$ for $c_j < 0$. This means that the potential method can make the perfect learning for given teachear's data without doing anything. If we use the potential method as it is, however, it yields several small isolated influence regions just like "islands" in many problems. Clearly, this phenomenon causes a poor generalization ability. Therefore, in oder to obtain as smooth a discriminant surface as possible, we adjust the charges of data. This is the learning of the potential method.

A way of learning in the potential method can be summarized as follows:

Step 0. At the beginning, all teacher's data have an equal amount of charge except for the difference in its sign (suppose that each data of the class $\mathscr{A}$ has a positive charge, while each data of the class $\mathscr{B}$ a negative charge).

Step 1. Consider the ith pattern x_i $(i = 1, \cdots, N)$. Examine whether it is categorized correctly or not on the basis of the sign of output of

$$\tilde{P}(x_i) = \sum_{j \neq i}^{N} \frac{c_j}{D(x_i, x_j)}.$$

If the pattern x_i is not categorized correctly, then add the index i to the set I_{error}. If I_{error} is empty, then stop the iteration. Otherwise go to the next step.

Step 2. Find the pattern x_p with the highest error, namely

$$|\tilde{P}(x_p)| = \max_{i \in I_{error}} |\tilde{P}(x_i)|.$$

Step 3. Find the pattern x_q in the other category than of x_p nearest to the pattern x_p. Change the charge c_j of pattern x_j $(j = 1, \cdots, N)$ in such a way that the potential at $x_m = (x_p + x_q)/2$ becomes zero. Namely, suppose that the new charge $c_j^{'}$ is given by

$$c_j^{'} = c_j \exp\big(\mp \tilde{P}(x_j)\gamma\big) \quad j = 1, \cdots, N, \tag{1}$$

where denoting $q_j = c_j/D(x_m, x_j)$, $(j = 1, \cdots, N)$, γ solves

$$\begin{aligned} q_1 \exp\big(-\tilde{P}(x_1)\gamma\big) + \cdots + q_{N'} \exp\big(-\tilde{P}(x_{N'})\gamma\big) + \\ + q_{N'+1} \exp\big(\tilde{P}(x_{N'+1})\gamma\big) + \cdots + q_N \exp\big(\tilde{P}(x_N)\gamma\big) = 0. \end{aligned} \tag{2}$$

Here, $x_1, \cdots, x_{N'}$ have positive charges, while $x_{N'+1}, \cdots, x_N$ negative charges. The sign $\mp$ in Eq. (1) means that $c_j > 0$ takes "$-$" and $c_j < 0$ "+".
Replace the charge of each pattern by the new one given by Eq. (1), and go to Step 1.

Remark 1. In changing charges, we focus our attention on a data whose position has the highest potential in the opposite category. It is possible to consider all data whose positions have potentials with the opposite sign. In this event, the equation to be solved becomes a system of several nonlinear equations. Although the authors examined several methods for solving the system of nonlinear equations, any technique have some difficulties, say, being trapped in local minima, no convergence sometimes, time consuming and so on. Although the above method based on the Eq. (2) produces just an approximate solution to our modification problem of charges, it shows good performance in our experiences.

Remark 2. The potential method belongs to a class of kernel methods for machine learning in which the approximate function is given by

$$f(x) = \sum_{j=1}^{n} K_j(x, x_j) y_j,$$

where $y_j = 1$ for $x_j \in \mathscr{A}$ and $y_j = -1$ for $x_j \in \mathscr{B}$. In addition, the kernel $K_j(x, x_j)$ is a symmetric function that usually (but not always) satisfies the following properties [8]:

(i)	$K(x,x') \geq 0$	nonnegative,
(ii)	$K(x,x') = K(\|\|x-x'\|\|)$	radially symmetric,
(iii)	$K(x,x) = \max$	takes on its maximum when $x = x'$,
(iv)	$\lim_{t\to\infty} K(t) = 0$	monotonically decreasing with $t = \|\|x-x'\|\|$.

The potential methods uses the kernel $K(x,x') = = \frac{c}{||x-x'||}$ $(c > 0)$. In this event, the above property (iii) should be interpreted in such a way that the kernel has an infinite maximum when $x = x'$. Although the infinity property is not desirable in many mathematical analysis, it has a positive meaning in pattern classification problems. For cases in which the kernel is infinite at a test pattern, the potential at the test pattern has the correct sign without any learning. The only problem is that the generalization ability without adjustment of "charge" is poor in general. Therefore, the learning in the potential method is to adjust "charge" in order to increase the generalization ability.

Remark 3. The potential method can be extended by using a generalized potential

$$P(x) = \sum_{j=1}^{n} \frac{c_j}{\{D(x, x_j)\}^r}.$$

As r becomes larger, the influence of the data nearest to the test pattern gets larger. In the case of $r \to \infty$, therefore, the potential method with a generalized potential becomes the same as the k-nearest neighbour method with $k = 1$.

3. Additional learning

We can show that the additional learning can be made easily by using the potential method. Let x_t be a data added newly to the existing teacher's data. The procedure of additional learning can be divided into 1) the case in which x_t is classified correctly by the present rule, and 2) the case in which x_t is misclassified by the present rule. The details are as follows:

Case 1. When the new data x_t is classified correctly by the present rule, find a data x_a closest to x_t but in the different category of x_t. In addition, find a data x_b closest to x_a but in the different category of x_a. Let x_{abm} be the middle point of x_a and x_b, i.e., $x_{abm} = (x_a + x_b)/2$. If the potential of x_{abm} has a different sign from that of x_t, then put the charge c_t on x_t in such a way that we have

$$P'(x_{abm}) := P(x_{abm}) + c_t/D(x_t, x_{abm}) = 0.$$

Namely, we put

$$c_t = -P(x_{abm}) \times D(x_t, x_{abm}).$$

However, if the potential of x_{abm} has the same sign as that of x_t, then we do not put any charge on x_t (i.e, $c_t = 0$). The purpose of consideration of the potential of x_{abm} is to check whether the discriminant surface can be made correctly by adding x_t. Also, by excluding unnecessary data from additional learning, the computation time can be made shortened.

Case 2. When the new data x_t is misclassified by the present rule, find a data x_a closest to x_t but in the different category from x_t. Let $x_{atm} = (x_t + x_a)/2$. Then put the charge c_t on x_t in such a way that we have

$$P'(x_{atm}) := P(x_{atm}) + c_t/D(x_t, x_{atm}) = 0.$$

Namely, we put

$$c_t = -P(x_{atm}) \times D(x_t, x_{atm}).$$

4. Forgetting

If we make only additional learning according as some new knowledge are added, the newly obtained rule becomes more and more complex. Clearly, this does not give us a good effect in generalization ability of the method. Rather, it seems that unnecessary (or, inappropriate) rule in the present situation should be excluded. Human beings seem to grow up in such an adaptive way. Therefore, we should introduce forgetting as well as additional learning in machine learning.

How to forget is a difficult problem in machine learning. Maybe, one way is to forget unimportant data. In this event, we have to consider the degree of importance of data. In the potential method, the degree of importance for each data is considered to be given by the value of kernel function $K_i(x,x_i) = \frac{c_i}{D(x,x_i)}$.

4.1. Passive forgetting

In many situations, it seems natural that the degree of importance of data reduces as the time passes. A method for forgetting may be given by

$$c'_f = c_f \exp(-\alpha t),$$

where t denotes the time elapsed, α – the coefficient of forgetting, c_f – the original charge, and c'_f – the charge after t-time passed. Additionaly, it is supposed that the data x_f is extracted from the set of teacher's data, if t is beyond a threshold (the forgetting period).

The above method for forgetting depends only on the time elapse. However, it seems more effective to forget more actively data which give bad influences to correct judgment. We call the way of forgetting depending on the time elapse "passive forgetting", whereas the one of forgetting data with bad influence actively "active forgetting". We shall discuss the way of active forgetting in more detail below.

4.2. Active forgetting

A key for active forgetting is to find data giving a bad influence to correct judgment. We call such data "obstacle data". One way for finding obstacle data is given as follows. Suppose that a test pattern x_t is misjudged by the potential method. Let I_F denote the set of data in the other category from x_t. Removing a data $x_i \in I_F$, judge the category of test data x_t on the basis of its potential. If the judgment is correct, the data x_i is considered an obstacle data. Find such an obstacle data by checking all data $x_i \in I_F$.

Several ways for forgetting obstacle data is possible. Two simple ways (methods) are discussed below.

Method 1. Constant rate of forgetting with respect to the distance. The importance of obstacle data (i.e., the value of kernel) is decreased by controlling only the charge regardless the distance between the obstacle data and the test pattern. Let c_f denote the charge of the obstacle data x_f. A modified charge c'_f is given, for example, by

$$c'_f = \alpha c_f.$$

Here, the rate of forgetting α takes a value from [0,1].

Method 2. Increasing rate of forgetting with respect to the distance. In many cases, as the distance between a data x_i and the test pattern x_t becomes smaller, the influence of the data x_i becomes larger. Therefore, it seems natural to increase the rate of forgetting as the distance between the obstacle data and the test pattern becomes smaller. In this event, the value of kernel is controlled directly. One example is given by

$$K'_i = \alpha\beta K_i,$$

where α takes a value from [0,1], and β is given by

$$\beta = \frac{2}{1+e^{-\theta D(x_i,x_t)}} - 1.$$

The parameter θ is determined mainly by experience.

5. Applications to stock portfolio problems

5.1. Single stock investment

Our problem is to judge whether a stock is to be purchased or not. Seven economic indices are taken into account. We have the data in the 119 periods in the past for which it is already known to be purchased or not. We made a test of discriminant ability of the potential method taking the first 50 data as the teacher's ones, and examined the ability of classification for the rest 69 data. Figure 1 compares the result without additional learning and the one with additional learning with/without forgetting. Flags represent misclassified data. It can be observed that the additional learning provides a good effect in classification, in particular, around the period of 80's.

5.2. Portfolio mix problems

Our problem here is to make a portfolio mix among 213 stocks in the market. As in the previous subsection, it is known in the past 50 periods (1987.1-1991.2) whether each stock is to be purchased or not. In this event, each stock is considered in terms of 10 economic indices. The return rate is given by

$$\text{return rate} = \\ = \frac{\text{the highest price during the anteceding 6 periods}}{\text{current price}} - 1.$$

We judge a stock to buy if the return rate is over a certain threshold. In the following simulation, we suppose this threshold is 0.2 (Fig. 2).

Table 1
Index of advantage over the market

	Potential method		RBF network		1-NN method	
	Top30% Getting	Free Getting	Top30% Getting	Free Getting	Top30% Getting	Free Getting
Initial learning only	1.49 (63.0)	1.46 (100.4)	1.59 (63.0)	1.23 (103.8)	0.85 (63.0)	1.30 (130.3)
Additional learning (without forgetting)	1.40 (63.0)	1.64 (49.0)	1.85 (63.0)	2.37 (37.3)	1.05 (63.0)	1.65 (57.7)
Additional learning (with passive forgetting only)	1.58 (63.0)	2.17 (25.9)	1.75 (63.0)	2.84 (15.4)	1.01 (63.0)	1.60 (52.6)
Additional learning (with active forgetting only)	5.39 (63.0)	13.04 (49.9)	1.84 (63.0)	4.09 23.5	— —	— —
Additional learning (with active & passive forgetting)	6.62 (63.0)	24.99 (46.8)	2.11 (63.0)	8.48 (10.9)	— —	— —

The average number of invested stocks is indicated with a bracket.

Table 2
Index of advantage over the market for various forgetting schedules
(Free Getting active and passive forgetting)

Forgetting rate	r									
α	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1.00	1.91 (49.1)	2.56 (49.4)	2.87 (49.1)	3.77 (48.4)	6.10 (47.6)	7.28 (46.1)	9.51 (45.3)	14.90 (45.3)	16.29 (43.7)	21.12 (45.4)
0.70	2.66 (49.2)	3.20 (48.9)	3.77 (49.0)	4.65 (47.7)	6.29 (46.4)	9.61 (45.7)	11.10 (45.3)	12.81 (45.3)	17.09 (43.7)	18.78 (45.5)
0.50	7.28 (46.1)	6.88 (45.9)	8.83 (45.6)	9.61 (45.7)	9.34 (45.8)	9.81 (44.8)	12.56 (44.7)	13.79 (44.3)	14.22 (44.0)	18.57 (44.5)
0.30	6.24 (46.4)	6.58 (45.8)	8.41 (45.2)	10.46 (44.1)	12.94 (44.1)	13.79 (44.3)	14.37 (44.1)	18.20 (44.5)	16.01 (44.7)	21.17 (45.7)
0.10	15.84 (43.5)	16.66 (44.4)	15.33 (44.2)	17.11 (44.7)	16.87 (44.0)	18.57 (44.5)	18.34 (45.3)	22.54 (45.8)	20.13 (46.3)	21.72 (47.8)
0.09	17.25 (44.5)	16.86 (45.9)	19.15 (44.8)	20.96 (45.3)	19.83 (45.5)	21.57 (45.2)	19.69 (45.5)	23.38 (45.9)	19.36 (47.2)	22.26 (48.2)
0.07	17.43 (44.9)	17.85 (43.7)	16.87 (44.5)	18.22 (44.6)	21.87 (45.1)	22.60 (45.3)	22.21 (45.6)	23.37 (46.2)	23.48 (46.7)	20.78 (48.3)
0.05	23.30 (46.8)	24.13 (46.8)	21.05 (47.0)	22.24 (47.1)	20.17 (46.6)	20.84 (46.5)	21.75 (47.2)	24.99 (46.8)	23.10 (47.9)	19.41 (48.6)
0.03	20.78 (47.1)	20.30 (47.6)	20.24 (47.6)	21.31 (47.4)	20.19 (47.6)	22.41 (47.5)	21.26 (48.1)	21.59 (48.0)	19.54 (48.3)	21.34 (48.5)
0.01	22.52 (47.5)	21.43 (48.1)	19.91 (48.2)	20.04 (48.3)	19.03 (48.4)	19.42 (48.5)	19.35 (48.7)	21.09 (48.5)	21.36 (49.0)	20.89 (49.2)

The number of average invested stocks is indicated with a bracket.

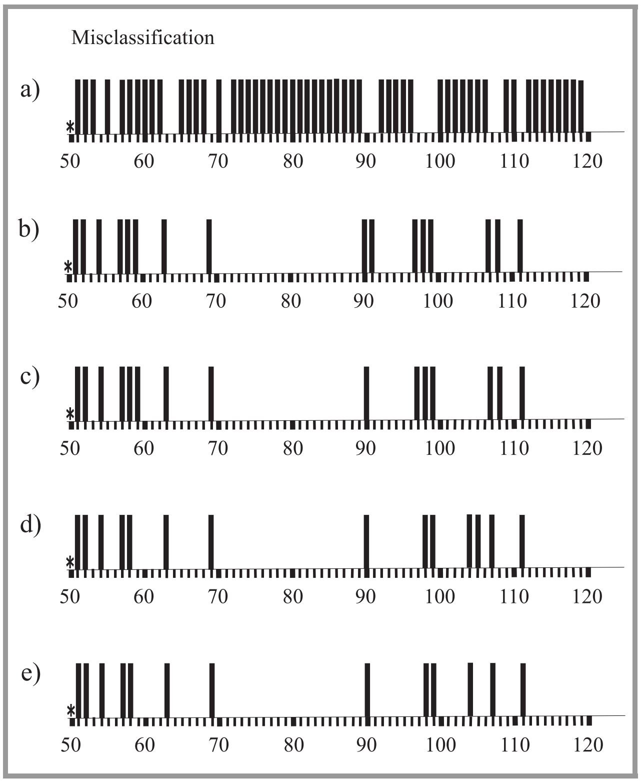


Fig. 1. Potential method with additional learning and forgetting: (a) initial learning, misclassified patterns 55; (b) only additional learning, misclassified patterns 16; (c) with passive forgetting, $r = -\lg 0.1/60$, misclassified patterns 15; (d) with active forgetting (method 1), $\alpha = 0.7$, misclassified patterns 14; (e) with active forgetting (method 2), $\alpha = 0.5$, misclassified patterns 13.

In the following, the way of Free Getting purchases only stocks which are judged to be purchased, while Top30% Getting the stocks of top 30% after sorting the stocks according to the degree of potential.

We made an examination of performance of portfolio mix by our method in the anteceding periods (1991.9-1996.3). Here, the index of advantage over the market I_A is defined by

$$I_A = \frac{(1+\alpha_1)(1+\alpha_2)\times\cdots\times(1+\alpha_T)}{(1+\beta_1)(1+\beta_2)\times\cdots\times(1+\beta_T)},$$

where α_i is the return rate of our portfolio mix at the ith period and β_i is the one of the market (usually called "index") at the ith period.

The initial learning was made for 50 periods between 1987.1 and 1991.2. The test with or without additional learning and forgetting is for 55 periods between 1991.9 and 1996.3.

The forgetting rates in cases with active forgetting only are $\alpha = 0.03$ and $r = 0.8$ for Top30% Getting, while $\alpha = 0.01$ and $r = 0.2$ for Free Getting. On the other hand, the forgetting rates in cases with active and passive forgetting $\alpha = 0.05$ and $r = 0.7$ for Top30% Getting, while $\alpha = 0.05$ and $r = 0.3$ for Free Getting. These values are the ones which provided the best result. Table 1 shows a comparison among potential method, RBF network and 1-*NN* method with various forgetting ways. The result for free getting with various forgetting schedules is shown in Table 2.

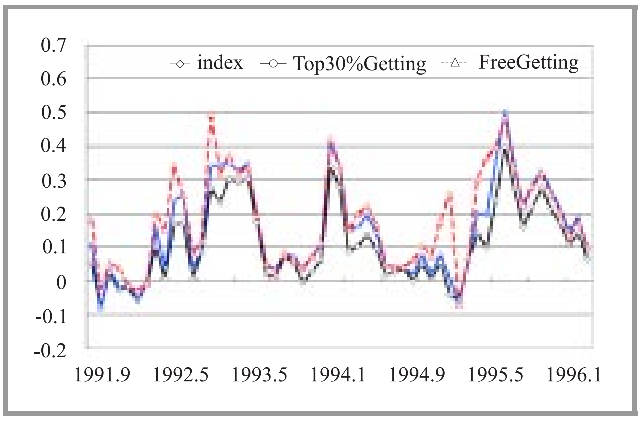


Fig. 2. Return rate by active and passive forgetting.

6. Concluding remarks

It has been observed that the effect of active forgetting is larger than that of passive forgetting. In general, the additional learning with forgetting provides a better performance than the mere additional learning. In the above example, however, the effect of appending forgetting to additional learning is not so remarkable in comparison with that of appending additional learning to the initial learning. In addition, the effectiveness of forgetting depends on its schedule. This implies that the forgetting is not so easy to use as the additional learning. It seems that human beings make forgetting in an effective way with almost optimal forgetting schedule on the basis of experience. Further examinations in practical problems are needed to find an optimal forgetting way in machine learning.

Acknowledgment

This work was supported by the Grant-in-Aid for Scientific Research of Education Ministry in Japan.

References

[1] H. Nakayama and N. Kagaku, "Pattern classification by linear goal programming and its extensions", *J. Glob. Opt.*, vol. 12, pp. 111–126, 1998.

[2] H. Nakayama and M. Yoshida, "Additional learning and forgetting by potential method for pattern classification", in *Proc. ICNN'97*, Houston, USA, 1997, pp. 1839–1844.

[3] H. Nakayama, "Growing learning mahines and their applications to portfolio problems", in *Proc. Int. ICSCC,* 1997.

[4] H. Nakayama, M. Yoshida, and S. Yanagiuchi, "Incremental learning for pattern classification", in *Proc. ICONIP'97*, Dunedin, New Zealand, 1997, pp. 498–501.

[5] H. Nakayama, S. Yanagiuchi, K. Furukawa, Y. Araki, S. Suzuki, and M. Nakata, "Additional learning and forgetting by RBF networks and its application to design of support structures in tunnel construction", in *Proc. Int. ICSC/IFAC Symp. Neural Comput. (NC'98)*, 1988, pp. 544–550.

[6] L. N. Cooper, "The hypersphere in pattern recognition", *Inform. Contr.*, vol. 5, pp. 324–346, 1962.

[7] D. L. Reilly, L. N. Cooper, and C. Elbaum, "A neural model for category learning", *Biol. Cybern.*, vol. 45, pp. 35–41, 1982.

[8] V. Cherskassky and F. Mulier, *Learning from Data: Concepts, Theory and Methods*. Wiley, 1998.

Hirotaka Nakayama
e-mail: nakayama@konan-u.ac.jp
Department of Information Science
and Systems Engineering
Konan University
8-9-1 Okamoto, Higashinada, Kobe 658, Japan

Kengo Yoshii
Department of Information Science
and Systems Engineering
Konan University
8-9-1 Okamoto, Higashinada, Kobe 658, Japan

Paper

A study on fractal dimensions and convergence in fuzzy control systems

Alfons Schuster, William Blackburn, and Miguel Segui Prieto

Abstract — This paper addresses a problem in the area of intelligent, knowledge-based systems, namely the generation of knowledge, by presenting a proposal for the automation of this task. The proposed approach is limited however by focusing on fuzzy control systems (FCSs). Results obtained from different experimental investigations indicate the potential of the approach.

Keywords — fuzzy control systems, chaos theory, convergence, fractal dimension.

1. Introduction

Knowledge is a central component in any intelligent, knowledge-based system. Problems obtaining or generating knowledge can arise from different sources. They may be due to the complexity of the domain, the accessibility and availability of domain knowledge or domain experts, the number of rules needed for a rule base, or the consistency and maintenance of such a base, for example [5]. Note that due to the focus of the paper the discussions here emphasise issues related to FCSs. Rules are not the only means by which knowledge is captured in a FCS. Fuzzy sets, their shape and arrangement, as well as the mechanisms by which they communicate in a system are also very important [15]. FCS design also very often has a strong trial and error nature in which the system designers very often play a vital role. One of the insights we gained from this underlying trial and error approach is that it is very often possible to generate multiple FCS solutions for the same problem. For example, the only difference between two FCS solutions could be the defuzzification technique they employ, but it also could be the slightly different shape of particular fuzzy sets. Another observation is that many FCSs show similarities in their dynamic behaviour. For example, the dynamic behaviour of a FCS application might look similar to the illustration given in Fig. 1.

This illustration actually is taken from an example application provided with the commercial fuzzy logic tool CubiCalc 2.0 that was used in this study. Note however that the circled line in the figure has been added manually to ease forthcoming discussions. Figure 1 illustrates the trajectories of two objects, A and B, moving from left to right in time. Y and X in the figure define a co-ordinate system. The objective of object B is to approach and finally catch object A. Both objects move with constant, but individual speeds, and so a dot or circle at position (X, Y) in the figure represents the position of an object in time. For simplicity object A moves on a straight line. Object B has to be more flexible due to the definition of its task. Note that although Fig. 1 illustrates two trajectories for object B, at the moment only the trajectory labelled with the number *1* is of interest. Figure 1 indicates that object B, following trajectory *1*, really approaches object A, and therefore provides a solution to the given task. It was mentioned earlier that this or a similar dynamic behaviour could be found in many other situations. Indeed, Fig. 1 could illustrate the movement of a robot arm trying to grasp an object on an assembly line, it could illustrate the control of the temperature in a room, but also the path of a remotely controlled vehicle on a planet approaching an object for probing, for example.

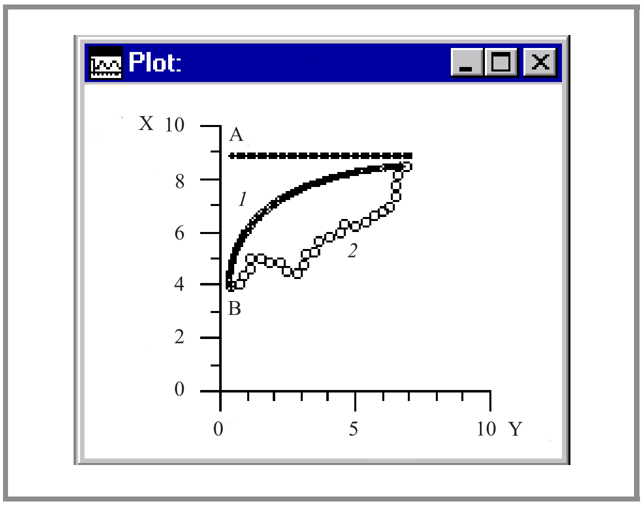


Fig. 1. Dynamic behaviour of an example FCS.

For later discussions it is also important to understand that with most commercial tools it is usually possible to record the values of selected variables (e.g., the X, Y positions of object A and B in Fig. 1) at each step in a time series. A time series therefore, in a sense, contains information about the dynamic behaviour of the system.

Another point that needs mentioning is illustrated by the second (circled) trajectory for object B in Fig. 1. Like trajectory *1*, this trajectory finally approaches object A, and thus, a FCS generating this trajectory could be regarded as a solution to the problem too. The solution finally selected however could be the FCS that produces trajectory *1*, because for many problems FCS designers prefer a system that converges towards a solution with some smoothness. Simply imagine the two trajectories in Fig. 1 as being proposed solutions for the robot arm mentioned before. It is not difficult to select the one more appropriate for the task. The following provides a synopsis of these observations:

- In the field of FCSs it is very often possible to generate multiple solutions for a problem.
- FCSs applications in different problem-solving situations show similarities in their dynamic behaviour.
- Convergence with a certain degree of smoothness can be a requirement in some FCS applications.

These observations form the basis for this work, which in broad terms can be summarised as a study investigating the similarities in FCSs, mentioned before. The means by which we aim to achieve this goal are:

1. The convergence of a proposed FCS solution is examined by a measure of convergence.
2. A fractal dimension algorithm on the other hand, determines the smoothness of a solution.

We investigated a number of FCSs and other models for quality assessment. The results generated in these studies indicate the potential of the approach. The reminder of the paper is organised as follows. Section 2 explains the measures we use in investigation in this study. Section 3 describes the FCSs and models we investigated. Section 4 presents the results from these investigations. Section 5 proposes an integrated system. Section 6 reviews related work, and Section 7 ends the paper with a summary.

2. Two measures used in this study

The measures introduced in this paper relate to some degree to what is sometimes loosely termed chaos theory. This theory has its origins in the study of nonlinear dynamical systems, and hence nonlinear differential equations [17]. It obtained increasing attention within the natural sciences about four decades ago. A key element being the fast progress in computer technology within this period [14]. With computes growing more and more powerful it was possible to investigate more and more complex systems with increasing efficiency. Lorenz, for example, investigated the extent to which weather is predictable [11]. Lorenz's work also produced an interesting by-product, the so-called Lorenz-attractors. The artistic beauty of many attractors led to an increasing awareness and popularity of the theory. Nowadays chaos theory is studied in many domains including medicine, engineering, and computer science, for example [3, 6, 8]. Out of these studies emerged a variety of new concepts and measures.

Before the two measures are explained in more detail we use the forthcoming section to discuss another concept from chaos theory that is important in the context of this paper, namely that of an attractor.

2.1. Attractors

Attractors, also often referred to as strange attractors, or fractals, are a very important concept in chaos theory. In chaos theory an attractor is more or less the state development of a dynamic system over time. The temporal development of these systems is often illustrated in so-called phase-state plots or phase diagrams. Very often these mathematically generated illustrations bear a striking similarity with structures we can find in nature. The shapes and forms of trees, lungs, shells, and clouds are typical examples [12]. To understand and connect mathematically generated attractors and fractals with these observations in nature is a strong motivation for the study of chaos theory, and so it is needless to say that a lot of work has been done in this area already. Little work however has been done on a particular view on attractors. In this particular view we suggest that the principle of an attractor appears quite frequently, often under different names, in our everyday life. The different expressions we use for the term attractor in many of these situations include the terms **goal**, **aim**, or **target**, for instance. The following two examples help illustrating this relationship. The goal of a person planning a holiday can be to be at a specific location over time. Or, the aim of an autonomous agent over time can be to avoid a number of obstacles. It is important to understand that the main objects (person, autonomous agent) in the two example systems move towards an **attractor**, or **goal**, or **aim** over time.

We humans are able to discuss natural structures, goals, and aims more or less elegantly via the use of our natural language. On the other hand, the language of chaos theory is mathematics. Although the mathematics of attractors and fractals can be relatively simple in some cases, it remains a fact that the mapping and interpretation of mathematical statements into the real world often can be very difficult, if not impossible. Let us therefore say:

– there is some sort of a gap between the mathematical world and the natural (problem-solving) world.

This observation makes this study in the area of FCSs interesting and promising, because by its very definition fuzzy logic provides a means for acting as a communicator between the mathematical world and the natural (linguistic) problem-solving world. Note also, that although the discussion here concentrates on attractors there are other concepts from chaos theory that are also very relevant in this context (e.g. self-similarity, and self-organisation) [16].

2.2. A measure of convergence

The previous section revealed that convergence could be an important feature in FCSs. For example, the task for object B in Fig. 1 was to approach and finally catch object A. The trajectories of the two objects consequently need to converge towards each other. The measure used in this study for distinguishing the convergence of different systems aims to reflect this behaviour. For example, let Fig. 2 illustrates the trajectories of two objects A and B.

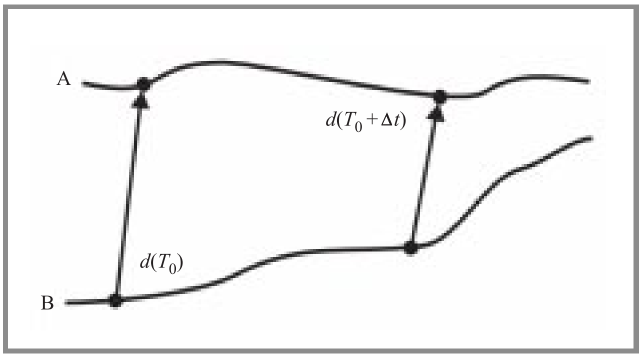


Fig. 2. Trajectories of two objects A and B.

In Fig. 2 $d(T_0)$ shall be the distance d between objects A and B at time T_0, and $d(T_0+\Delta t)$ the distance at time $t = T_0 + \Delta t$. In order to use the distance development between the two objects over time we here define a measure of convergence (*MOC*) as follows:

$$MOC = \frac{1}{N-1}\sum_{n=1}^{N-1} \lg\left|\frac{d_{n+1}}{d_n}\right| . \tag{1}$$

Note that the variable N in the equation stands for the number of data points in the time series. The features of this measure that could be useful in this study are:

- $MOC < 0$, may be an indicator for a system that produces convergent trajectories.
- $MOC = 0$, might indicate a system that is in some sort of steady state mode, for example, objects A and B moving on two parallel lines.
- $MOC > 0$, very likely an indicator for a system that produces non-convergent trajectories.

Here it could be interesting to refer to the so-called Lyapunov Exponent λ found in chaos theory. The Lyapunov Exponent is a measure to assist in the distinguishing between different types of orbits or trajectories of dynamic systems [10]. It is based on the mean exponential rate of divergence of two initially close trajectories, and describes the dynamic of a system qualitatively as:

- $\lambda < 0$, the orbit is attracted to a stable fixed point or a stable periodic orbit.
- $\lambda = 0$, the orbit is a neutral fixed point. The system is in some sort of steady state mode, like a satellite in a stable orbit, for example.
- $\lambda > 0$, the orbit is unstable and chaotic. Nearby points, no matter how close diverge to any arbitrary separation.

Although it is not the intention here to use the *MOC* for determining whether a system is chaotic or not it is interesting here to identify the similarity it bears with the Lyapunov Exponent.

2.3. A measure of smoothness

Section 1 suggested that convergence alone is very frequently not the only criterion when developing FCSs. Very often a solution should have certain smoothness too. The basic assumption is that, given different FCS solutions, a smoother trajectory is more likely to be selected than a trajectory that is rather jagged or irregular.

The study of so-called fractals may provide a possibility for quantifying the shape of a trajectory in terms of its smoothness, or jaggedness, respectively. Very generally, fractals are patterns or structures which, when being dealt with mathematically, produce results or properties that are difficult to be interpreted, or conflicting with predictions of traditional mathematics. An example would be the Koch-snowflake curve, a geometric object with finite area, but infinite circumference. Outstanding mathematicians attempted to come to grips with these objects. Mandelbrot for example, associates these pathological structures with forms that can be found in nature [12]. Hausdorff and Besicovitch on the other hand came forward with a general definition for the calculation of a (fractal) dimension for such objects. Their definition of a fractal dimension is based on an investigation of how geometric figures fill the space in which they are represented [4]. It is important here to mention that there exist many definitions for measurements on fractals. This paper, for instance, uses a method proposed by Gough for the calculation of a fractal dimension [7]. Also, remember that the geometric objects investigated here are time series representing the development of the distance between the trajectories of two objects. However, let the time series illustrated in Fig. 3 represents the distance development of an example system.

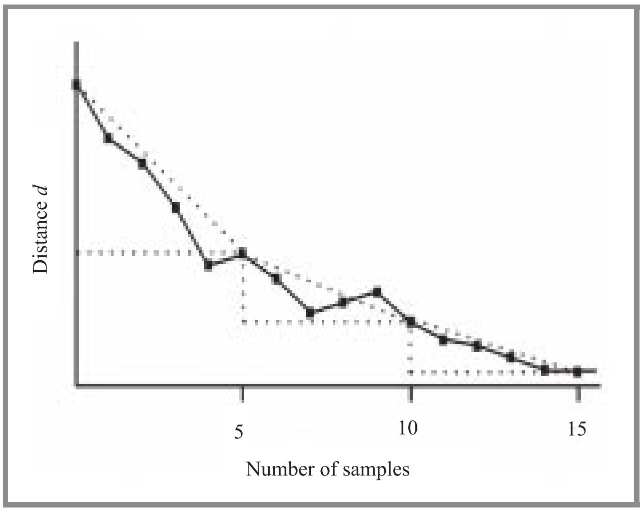


Fig. 3. Distance development of two trajectories, and length estimation using a ruler length of five.

Figure 3 illustrates that individual distance measurements are connected to a continuous line. Gough's method is used to calculate a fractal dimension from such a line. Initially the method determines different estimates of the length L of the line by measuring it with different so-called rulers of length r. The line in Fig. 3 for instance is measured with a ruler of length five. A length measurement for a particular ruler is determined by the following equation:

$$MOC = \frac{1}{N-1}\sum_{i=1}^{N-1}\left[\left\{r^2+(x_{ir}-x_{(i+1)r})^2\right\}^{\frac{1}{2}}+\right.$$
$$\left.+\left\{(N-rk-1)^2+(x_{rk}-x_{N-1})^2\right\}^{\frac{1}{2}}\right].$$

In this equation $k = \mathrm{Trunc}\left[\frac{\mathrm{N-1}}{\mathrm{r}}\right]$, r represents the ruler length, and N the number of distance measurements in the time series. In simple terms a single length estimate (L_r) is a summation of hypotenuses. In order to extract a fractal dimension from such a diagram the method then plots the logarithm of the length estimates ($\lg L_r$) against the logarithm of the ruler length ($\lg r$). Figure 4 illustrates an example of such a graph.

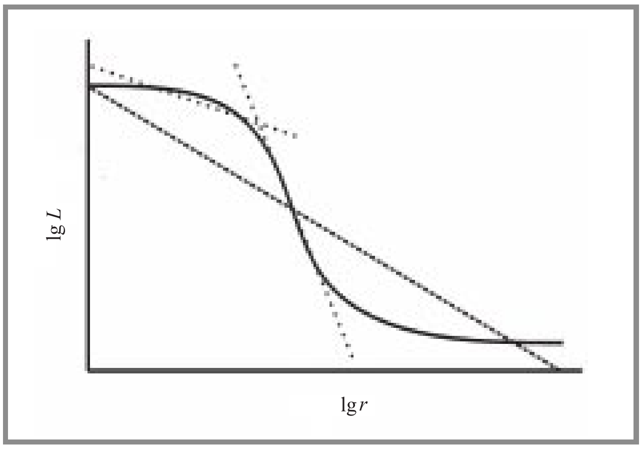


Fig. 4. Extraction of a fractal dimension.

The establishment of a fractal dimension from such a diagram is not that simple however. The traditional definition by Hausdorff and Besicovitch leads towards using the slope of a regression line (dashed line in Fig. 4) through the data points as an approximation for a fractal dimension. Other researchers came up with other interpretations. Kaye for example generates regression lines and fractal dimensions for separate regions in a plot (the two dotted lines in Fig. 4 for example), and compares these fractal dimensions with the features of "structure" and "texture" in fine-particle science [9]. This paper follows Kaye's view, and so it could be said that a measurement with longer rulers identifies the global behaviour (structure) of the distance development between two trajectories. On the other hand, measurement with smaller rulers provides information about the behaviour of the distance function at smaller scales (texture).

3. Investigated systems

Figure 5 illustrates some of the systems we investigated. The systems will be referred to as System 1, 2, 3, and so forth. The first three systems are FCS applications taken from an example library that is included in the software tool that has been used in the study. System 1 has already been introduced in Section 1 and therefore a description of it is omitted here. System 2 is a FCS that controls the movement of a truck (B) that tries to enter a parking slot. Figure 5 illustrates three parking attempts. The starting position of the truck is always randomly selected. The parking slot (A) remains at position 50.0 on the x-axis. The three scenarios in Fig. 5 show that the trajectories produced by the truck always converge towards the parking slot. FCS System 3 faces the problem of trying to suspend a metal object (B) in air at a stable position midway between an electromagnet at height 10.0 and the ground (height 0.0). Figure 5 illustrates two attempts. For example, take the attempt where the initial position of the metal object is at height 7.0 between the ground and the electromagnet. The FCS controls the magnetic field generated by the electromagnet according to the position of the metal object between the magnet and the ground. The field is continually changed until object (B) is suspended midway (height 5.0) between the electromagnet and the ground. This position is labelled (A) in Fig. 5. The x-axis in the figure represents the number of iterations the FCS goes through over time. Figure 5 illustrates that the trajectories produced in both attempts represent a solution to the problem.

To make the study more comprehensive we investigated various other systems. Some of these systems, Systems 4 to System 12, are illustrated in Fig. 5. Each illustration in Fig. 5 contains two trajectories (y_1 and y_2), corresponding to the movement of two imaginary objects. For example, the first trajectory for System 4 is defined by the exponential function $y_1 = 100\,\mathrm{e}^{-0.02x}$, and the second trajectory by the function $y_2 = 0$. Note that apart from System 6 and System 9 the second trajectory is always defined as $y_2 = 0$. Note also that the range for the values along the x-axis is the same for these system, namely [0, 200]. Further, System 4, 5, and 6 illustrate systems exhibiting convergent behaviour, whereas System 7, 8, 9, 10, 11, and 12 are used to represent non-convergent behaviour. The non-convergent systems can be further divided. System 10 and 11 indicate objects moving along in parallel (System 10) or oscillating parallel (System 11). Trajectory y_1 of System 12 finally was generated randomly.

4. Results

Table 1 illustrates *MOCs* and fractal dimensions extracted from these systems using the techniques described before.

Column 1 in Table 1 indicates the system, and column 2 the number of data points in a time series. Column 3 holds the *MOC* for each system. Column 4 and 5 finally contain fractal dimensions. The two columns differ in using different sets of rulers for the measurement of a fractal dimension of a time series. For example, taking a system with 200 data points, Ruler 1 to 10 means that the time series has been measured with rulers of length 1, 2, . . . , 10. For the same

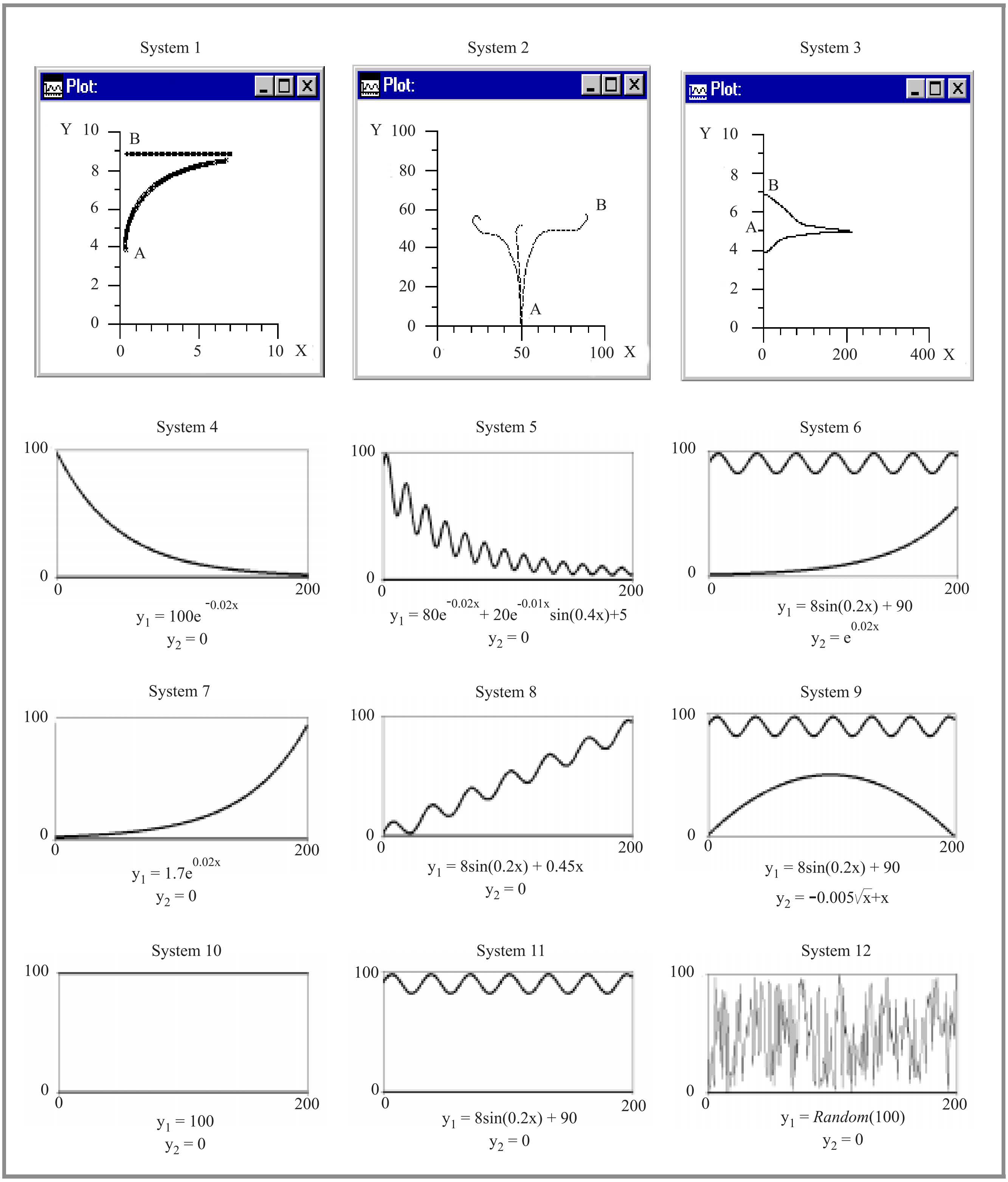


Fig. 5. Example systems studied in this work.

Table 1
MOCs and fractal dimensions of some of the example systems investigated in this study

System	Data points	*MOC*	Ruler (1 to 10)	Ruler (10 to 40)
System 1	75	-0.0280	1.0000	1.0000
System 2	61	-0.0342	1.0000	1.0010
System 3	200	-0.0070	1.0000	1.0000
System 4	200	-0.0086	1.0000	1.0021
System 5	200	-0.0069	1.0003	1.0022
System 6	200	-0.0017	1.0355	1.2250
System 7	200	0.0086	1.0001	1.0019
System 8	200	0.0084	1.0318	1.1776
System 9	200	0.0001	1.0355	1.2250
System 10	200	0.0000	1.0000	1.0000
System 11	200	0.0001	1.0355	1.2250
System 12	200	-0.0052	1.9754	1.6497
	200	0.0010	1.9510	1.6183

200 data points, Ruler 10 to 40 stands for a measurement with rulers of length 10, 11, …, 40.

4.1. Discussion of results

Table 1 illustrates that the *MOCs* of the six convergent systems (System 1, 2, 3, 4, 5, and 6 in Fig. 5) are all negative. The *MOCs* of the non-convergent systems (System 7, 8, and 9) are all positive. The *MOC* of System 10 (parallel) is zero, and that of System 11 (oscillating parallel) is very close to zero. These results are encouraging, because the *MOC* so far separates convergent from non-convergent systems. They are also interesting when being compared with the qualitative interpretation of a Lyapunov Exponent in Section 2.1, where a negative exponent indicated stable systems, a positive exponent unstable systems, and one of zero systems that are in some sort of steady state. Table 1 however also reveals that it is possible to obtain positive as well as negative *MOCs* for different random systems (System 12). Initially this seems to be problematic, but the fractal dimension values in column 4 and 5 indicate a possible solution to this problem. Remember that the "preferred" solutions are less jagged and irregular, and so should have a smaller fractal dimension. The fractal dimension values for the two random examples clearly reflect this assumption. It is also interesting to see that the three FCSs, as well as System 4, 7, and 10 all have very low fractal dimensions (close to 1.000), which is corresponding to the smoothness they illustrate. The remaining systems, apart from System 5, all have higher fractal dimensions. Note that although this discussion refers to the values in column 4 in Table 1, an interpretation of column 5 leads to similar observation.

It was mentioned earlier that the systems in Fig. 5 are representative instances of a larger group of systems we investigated. For example, the *y*-axis for System 4 to System 12 in Fig. 5 is scaled from 0 to 100 in this paper, but we also have evaluated systems showing similar trajectories at different scales. The results established by these other systems did allow an interpretation similar to the interpretation given before. From the viewpoint of the motivation behind this paper the results established in this study therefore can be interpreted as quite positive and encouraging to undertake further research in this direction.

5. Proposal for an implementation

Figure 6 at the end of this section illustrates our vision of a system that could be capable to automatically generate components for the knowledge base of a FCS application.

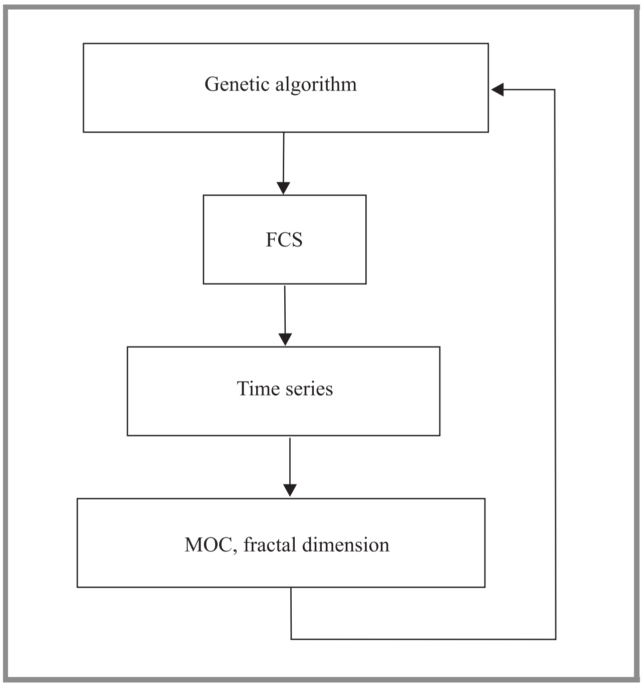


Fig. 6. Proposal for an implementation of the techniques presented before into a full system.

Figure 6 basically illustrates the integration of the methods presented in this paper with a genetic algorithm (GA). For example, the GA initially generates a pre-defined number of FCSs. Each of these FCSs is tested, and each of them produces a time series when tested. On the basis of this time series it is possible to estimate the potential of a FCS according to the *MOC* and the fractal dimension it produces. Solutions that indicate as being better than others are selected and modified by the GA to achieve further improvement. This process runs until a pre-defined threshold is reached. A system developer would evaluate the final proposal of the system.

Certainly, this process can be implemented at different levels of complexity. The GA could be used for the generation of a rule base only. Additionally it could be used for the generation of fuzzy sets, and the selection of different inference mechanisms. Our intention therefore is to begin with the testing of less complex systems. This strategy is also supported by the fact that very often the description of a problem and its solution could be very simple in a FCS. For example, FCS System 1 uses only five rules, eight fuzzy sets, one input variable, and one output variable.

6. Related work

Control systems, including FCSs have been studied extensively, with different interests, in the past [1, 15]. This section mentions some of the work that motivated us in our research.

Chen and Hwang for example, indicate that it is nearly always possible to describe FCS applications in completely different domains with a relatively small number (about five to eight) of very often similar fuzzy sets [2]. An early paper by Miller supports Chen and Hwang's work by identifying the number seven plus/minus two as a benchmark in many complex situations [13]. For example, instead of a lengthy explanation chess player often only mention a small number of key features of a game. These examples so far parallel the observations mentioned earlier here in terms of the simplicity and the similarity of many FCS applications. The simplicity aspect in particular can be advantageous for the system we bear in mind. For example, the discussion so far suggest the generation of a relatively small number of fuzzy sets for a FCS by the GA in Fig. 6, and this would keep the complexity of the full system low. Further relevant material can be found in a paper by Schuster [16]. Schuster discusses the relationship between self-similarity in chaos theory and so-called adaptive fuzzy sets in the context of intelligent systems. Schuster argues that a set of fuzzy sets used for the description of a system variable often can be used for the same variable at different scales, but also very often for a completely different variable. Finally, in the field of FCSs researchers nearly always emphasise the trial and error nature of the development process and the importance of the system developer. The integration of the techniques presented in this paper with a search strategy such as a genetic algorithm therefore seems to be very promising for the problem at hand.

7. Summary

This paper presented a proposal for the task of automated knowledge generation. The proposal includes ideas and concepts from chaos theory. The results we obtained from different experimental investigations are encouraging in our opinion. Although our study concentrated on a particular type of knowledge based systems, namely FCSs, we belief that the presented approach may have the potential to be useful for a wider range of problems. Our current efforts revolve around an implementation of the presented proposal in a system similar to the system described in Section 5.

Acknowledgements

We want to thank Nortel Networks and IRTU for their support in this work.

References

[1] G. A. F. Aly, M. N. Aly, A. A. Shoukry, and A. A. Daby, "Different techniques for tracking non-linear control systems.", in *CDROM 6th Int. Conf. Inform. Syst., Anal. Synth. ISAS '2000*, Orlando, USA, 2000.

[2] S. J. Chen and C. L. Hwang, *Fuzzy Multiple Attribute Decision Making, Methods and Applications*. Berlin, Heidelberg: Springer Verlag, 1992.

[3] N. Crook and C. Dobbny, "Identifying trajectories in dynamic systems", in *Proc. Int. ICSC Symp. Eng. Intel. Syst.*, Tenerife, Spain, 1998, pp. 469–475.

[4] G. Elert, "The chaos hypertextbook",
http://www.hypertextbook.com/chaos

[5] J. Finlay and A. Dix, *An Introduction to Artificial Intelligence*. UCL Press, 1996.

[6] A. L. Goldberger, "Fractal mechanisms in the electrophysiology of the heart", *IEEE Eng. Med. Biol.*, pp. 47–51, 1992.

[7] N. A. J. Gough, "Fractal analysis of foetal heart rate variability", *Phys. Meas.*, vol. 14, pp. 309–315, 1993.

[8] P. Grim, "Self-reference and chaos in fuzzy logic", *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 4, pp. 237–253, 1993.

[9] B. H. Kaye, *A Random Walk Through Fractal Dimensions*. Ed. Weinheim, Cambridge, New York, 1989.

[10] H. J. Korsch and H. J. Jodl, *A Program Collection for the PC*. Berlin, Heidelberg: Springer Verlag, 1994.

[11] E. N. Lorenz, "Deterministic non-periodic flow", *J. Atmos. Sc.*, vol. 20, p. 130, 1963.

[12] B. B. Mandelbrot, *The Fractal Geometry of Nature*. New York: Freeman & Company, 1977.

[13] G. A. Miller, "The magic number seven, plus or minus two", *Appl. Phys. B*, vol. 63, pp. 81–97, 1965.

[14] T. S. Parker and L. O. Chua, "Chaos: a tutorial for engineers", *Appl. Phys. B*, vol. 75, no. 8, pp. 982–1008, 1987.

[15] T. J. Ross, *Fuzzy Logic with Engineering Applications*. New York: McGraw-Hill, 1995.

[16] A. Schuster, K. Adamson, and D. A. Bell, "Problem-solving in a self-similar world and adaptive fuzzy sets", in *Proc. IASTED Int. Conf. Artif. Intel. Soft Comput.*, Honolulu, Hawaii, USA, 1999, pp. 193–196.

[17] I. Stewart, *Does God Play Dice? The New Mathematics of Chaos*. London: Penguin, 1997.

Alfons Schuster
e-mail: a.schuster@ulst.ac.uk
Faculty of Informatics
School of Information and Software Engineering
University of Ulster
Shore Road, Newtownabbey, Co. Antrim, Northern Ireland

Paper

Multi-agent utility theory for ethical conflict resolution

Hiroyuki Tamura

Abstract — In this paper we try to construct a two-attribute group disutility function for two conflicting decision makers, taking into account the property of utility independence and/or convex dependence between them. Two variables in the group utility function are disutility levels of two conflicting decision makers. The disutility level of each decision maker is modeled using multiple attributes, that is disutility function of each decision maker is formulated as a multi-attribute disutility function. By using a group disutility function for two conflicting agents, we can model the mutual concessions of the two conflicting agents taking into account ethical preference of each decision maker, and hence we can expect fairer multiple agents decision making for realizing better social welfare.

Keywords — utility theory, conflict resolution, multi-agent decision making, ethical consideration, convex dependence.

1. Introduction

In planning a large public project it is quite important to obtain a consensus between two conflicting decision makers, a representative of the regional inhabitants and of the enterpriser, to create a pleasant and useful environment. This paper deals with a methodology of modeling decision analysis for consensus formation between two conflicting multiple agents, regional inhabitants and the enterpriser (developer) of a big public project such as constructing a large international airport, a freeway with heavy traffic, a refuse incineration plant, etc. For this purpose we try to construct a group disutility function for two conflicting agents, taking into account the utility independence [1, 1993] or convex dependence [2] between them. This is called the "multi-agent utility theory". By using such a group disutility function for two conflicting agents, we can model the mutual concessions of the two conflicting agents taking into account ethical preference [3] with each other, and hence we can expect fairer MADA (multiple agents decision making) for realizing better social welfare.

2. A group disutility functions for multi-agent decision making

Let $D_1 \times D_2$ be a two-attribute disutility function space and $d_1(x_1) \in D_1$, $d_2(x_2) \in D_2$ denote the disutility functions of decision maker DM1 and DM2 on the multi-attribute consequence spaces X_1 and X_2, respectively, where $x_i \in X_i$ $(i = 1, 2)$ denotes a specific consequence for DM*i*.

For a given $d_1(x_1) \in D_1$, and $d_2(x_2) \in D_2$, a group disutility function on $D_1 \times D_2$ space is defined as

$$G(x_1, x_2) = g\big(d_1(x_1), d_2(x_2)\big) \equiv g(d_1, d_2)\,.$$

Let us assume that d_1^0 and d_2^0 denote the worst levels of disutilities of DM1 and DM2, respectively, and d_1^* and d_2^* denote the best levels of disutilities of DM1 and DM2, respectively. Given an arbitrary $d_2 \in D_2$, a normalized conditional group disutility function (NCGDF) of DM1 is defined as

$$g_1(d_1 \mid d_2) \equiv \frac{g(d_1, d_2) - g(d_1^*, d_2)}{g(d_1^0, d_2) - g(d_1^*, d_2)}\,,$$

where it is assumed that

$$g(d_1^0, d_2) > g(d_1^*, d_2)\,.$$

It is obvious that

$$g_1(d_1^0 \mid d_2) = 1, \quad g_1(d_1^* \mid d_2) = 0$$

that is, NCGDF is normalized and is a *single-attribute* group disutility function. Hence it is easily identified.
The NCGDF for DM2, that is, $g_2(d_2 \mid d_1)$ can also be defined similarly as

$$g_2(d_2 \mid d_1) \equiv \frac{g(d_1, d_2) - g(d_1, d_2^*)}{g(d_1, d_2^0) - g(d_1, d_2^*)}\,.$$

The NCGDF $g_1(d_1 \mid d_2)$ represents DM1's and $g_2(d_2 \mid d_1)$ represents DM2's subjective preference for the group disutility as a function of his own disutility level, under the condition that the disutility level of the other DM is given. If NCGDF $g_1(d_1 \mid d_2)$ does not depend on the conditional level d_2, then attribute D_1 is utility independent [1, 1993] on attribute D_2. If attributes D_1 and D_2 are mutually utility independent, the two-attribute disutility function $g(d_1, d_2)$ can be described as either a multiplicative or additive from [1, 1993].
Suppose

$$g_1(d_1 \mid d_2) \neq g_1(d_1 \mid d_2^*)$$

for some

$$d_2 \in D_2$$

that is, utility independence does not hold between two attributes D_1 and D_2. In this case we can use a property of convex dependence [2] as a natural extension of utility independence.

The property of convex dependence is defined as follows: attribute D_1 is mth order convex dependent on attribute D_2, denoted $D_1(\mathrm{CD}_m)D_2$, if there exist distinct $d_2^0, d_2^1, \ldots, d_2^m \in D_2$ and real functions $\lambda_0, \lambda_1, \ldots, \lambda_m$ on D_2 such that NCGDF $g_1(d_1 \mid d_2)$ can be written as

$$g_1(d_1 \mid d_2) = \sum_{i=0}^{m} \lambda_i(d_2)\, g_1(d_1 \mid d_2^i),$$

where

$$\sum_{i=0}^{m} \lambda_i(d_2) = 1$$

for all $d_1 \in D_1$ and $d_2 \in D_2$ where m is the smallest non-negative integer for which this relation holds.

This definition says that, if $D_1(\mathrm{CD}_m)D_2$, then any NCGDF on D_1 can be described as a convex combination of $(m+1)$ NCGDFs with different conditional levels where $\lambda_i(d_2)$s are not necessarily non-negative. Especially, when $m = 0$ and $D_1(\mathrm{CD}_0)D_2$, attribute D_1 is utility independent on attribute D_2.

The algorithm for constructing a two-attribute group disutility function is as follows:

Step 1. NCGDFs $g_1(d_1 \mid d_2^0)$, $g_1(d_1 \mid d_2^*)$ and $g_1(d_1 \mid d_2^{0.5})$ are assessed, where $d_2^{0.5}$ denotes the intermediate level of attribute D_2 between the worst level d_2^0 and the best level d_2^*.

Step 2. If these NCGDFs are almost identical, $D_1(\mathrm{CD}_0)D_2$ holds. Otherwise, go to Step 3.

Step 3. If the convex combination of $g_1(d_1 \mid d_2^0)$ and $g_1(d_1 \mid d_2^*)$ is almost identical with $g_1(d_1 \mid d_2^{0.5})$, $D_1(\mathrm{CD}_1)D_2$ holds. Otherwise, higher order convex dependence holds. Once the order of convex dependence is found, the decomposition form [2] two-attribute disutility function can be obtained. Single-attribute NCGDFs play a role of basic elements in the two-attribute group disutility function.

Step 4. By assessing the corner values of a group disutility function in two-attribute space, coefficients of linear terms in the two-attribute group disutility function are obtained [4]. As a result a two-attribute group disutility function is obtained.

In modeling multi-agent decision making with conflicting DMs, NCGDF plays the most important role as it can model various patterns of a DM's preference who is self-centered and selfish or flexible and cooperative, and so forth.

3. Consensus formation modeling for multi-agent decision making

Let DM1 and DM2 be

- DM1: representative of the regional inhabitants;
- DM2: representative of the enterpriser who is planning a new public project.

Suppose the disutility function d_1 for DM1 evaluates environmental impact from the public project and the disutility function d_2 for DM2 evaluates the cost to realize various countermeasures of the public project. These disutility functions are constructed by questioning the environmental specialists about each situation of DM1 and DM2.

We construct the NCGDFs by again questioning the environmental specialists about each situation of DM1 and DM2. Consequently, suppose we obtained three types of models as follows:

Model 1. Mutual utility independence holds. Both DM1 and DM2 do not think that group disutility is small unless their own disutility is also small. In this case both DM1 and DM2 are selfish and strongly insist upon their own opinion. This situation shows the initial phase of planning a new project, when the plan has just been presented to the regional inhabitants.

Model 2. Utility independence holds for DM1 and first order convex independence holds for DM2. The attitude of DM1 is almost the same as in Model 1, however, DM2 is becoming more flexible towards obtaining consensus of DM1. In this case DM1 does not have enough information on the project, however, DM2 has obtained various information. This situation corresponds to the second phase of the consensus formation process.

Model 3. Mutual first order convex independence holds. The attitude of both DM1 and DM2 is getting more flexible and cooperative. In this case both DMs have obtained sufficient information about planning the public project and the countermeasures for preventing environmental impacts from the project, and thus, show a mutual concession taking into account ethical consideration with each other. This situation corresponds to the final phase of the consensus formation process between DM1 and DM2.

Suppose the minimum value of group disutility is obtained for Model 3. This implies that the most impartial consensus formation is obtained under the situation of Model 3, which is based on convex dependence between two conflicting DMs.

As seen from the consensus formation model described above it may be used as a fundamental material for discussion when the regional inhabitants and the enterpriser of a public project regulate and adjust their opinion of each other.

4. Concluding remarks

By using a group disutility function for two conflicting agents, we could model the mutual concessions of the two conflicting agents taking into account ethical preference of each decision maker. We believe that the group disutility function proposed in this paper is the first mathematical model that can handle ethical preference of conflicting decision makers with each other. The key idea of this mathematical model is that the two-attribute group disutility function is a function of single-attribute normalized conditional group disutility function of each decision maker. Ethical preference of each decision maker is described in this NCGDF.

The consensus formation model described in this paper is expected to be used as a fundamental material for discussion when the enterpriser of a public project and the regional inhabitants regulate and adjust their opinion with other for realizing better social welfare.

References

[1] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. 1993 (originally publised New York: Wiley, 1976).

[2] H. Tamura and Y. Nakamura, "Decompositions of multiattribute utility functions based on convex dependence", *Oper. Res.*, vol. 31, no. 3, pp. 488–506.

[3] A. Sen, *Collective Choice and Social Welfare*. Amsterdam: North-Holland, 1979.

[4] H. Tamura and T. Yoshikawa, Eds., *Large-Scale Systems Control and Decision Making*. New York: Marcel Dekker, 1990.

Hiroyuki Tamura
e-mail: tamura@sys.es.osaka-u.ac.jp
Department of Systems and Human Science
Graduate School of Engineering Science
Osaka University
1-3 Machikaneyama, Toyonaka
Osaka 560-8531, Japan

Paper

On equitable approaches to resource allocation problems: the conditional minimax solutions

Włodzimierz Ogryczak and Tomasz Śliwiński

Abstract — In this paper we introduce and analyze a solution concept of the conditional minimax as a generalization of the minimax solution concept extended to take into account the number of services (the portion of demand) related to the worst performances. Namely, for a specified portion of demand we take into account the corresponding portion of the maximum results and we consider their average as the worst conditional mean to be minimized. We show that, similar to the standard minimax approach, the minimization of the worst conditional mean can be defined by a linear objective and a number of auxiliary linear inequalities. We report some results of initial computational experience with the new solution concept.

Keywords — telecommunication networks, resource allocation, equity, min-max.

1. Introduction

Resource allocation problems are concerned with the allocation of limited resources among competing activities so as to achieve the best overall performances. In this paper, we focus on approaches that, while allocating resources, attempt to provide an equal treatment of all the competing activities [8]. The problems of efficient and equitable resource allocation arise in various systems which serve many users, like in telecommunication systems among others. Telecommunication networks are expected to satisfy the increasing demand for traditional services as well as to accommodate multimedia services. Hence, it becomes critical to allocate network resources, such as available bandwidth, so as to provide high level performance of all services at numerous destination nodes. The performance can be measured in terms of expected delays to be equitably minimized for all service demands.

The generic resource allocation problem may be stated as follows. Each activity is measured by an individual performance function that depends on the corresponding resource level assigned to that activity. A smaller function value is considered better, like the performance measured in terms of expected delays. Models with an (aggregated) objective function that minimizes the mean (or simply the sum) of individual performances are widely used to formulate resource allocation problems, thus defining the so-called minisum solution concept. This solution concept is primarily concerned with the overall system efficiency. As based on averaging, it often provides solution where low demand services are discriminated in terms of delays. An alternative approach depends on the so-called minimax solution concept, where the worst performance (maximum delay) is minimized. The minimax approach is consistent with Rawlsian [11] theory of justice, especially when additionally regularized with the lexicographic order [9]. On the other hand, allocating the resources to optimize the worst performances may cause a large worsening of the overall (mean) performances.

In this paper we introduce and analyze an alternative compromise solution concept of the conditional minimax. It is a parametric generalization of the minimax solution concept taking into account the number of services (the portion of demand) related to the worst performances. Namely, for a specified tolerance level (number of services k or portion of demand β) we take into account the entire group of the k (β portion) maximum results and we consider their average as the worst conditional mean to be minimized. According to this definition the solution concept is based on averaging restricted to the group of the worst results. We show that, similar to the standard minimax approach, the minimization of the worst conditional mean can be defined by a linear objective and a number of auxiliary linear inequalities.

Resource allocation models may be used to help to solve two major types of telecommunication problems emerging with exploding demand on multimedia services [2]. The first type of problems is related to decision support for designing robust and cost-effective fiber-optic networks [3]. The other field is traffic engineering which represents the ability to optimize the use of network resources only by means of efficient routing decisions [4]. In other words, while the first group of problems deals with the network engineering being related to the physical design of the network, the second group is rather related to the software design. The proposed solution approach is general enough to be applicable for both types of problems. However, we demonstrate it on straightforward problems related to the traffic engineering.

The paper is organized as follows. In the next section we introduce our generic resource allocation model and we show how it can be used to express several traffic engineering problems. In Section 3 the solution concept of the conditional minimax is formally introduced and it is shown that, similar to the standard minimax approach, the solution can be defined by a linear objective and a number of auxiliary linear inequalities. In Section 4 we report some results of our initial computational experience with the new solution concept.

2. The model

The generic resource allocation problem that we consider may be stated as follows. There is given a set of m services. There is also given a set Q of allocation patterns (allocation decisions). For each service i a function $f_i(\mathbf{x})$ of the allocation pattern $\mathbf{x}$ has been defined. This function, called the individual objective function, measures the outcome (effect) $y_i = f_i(\mathbf{x})$ of the allocation pattern for service i. In applications, we consider, an outcome usually expresses the delay. However, we emphasize to the reader that we do not restrict our considerations to the case of outcomes measured as delays. They can be measured (modeled) as service time, service costs as well as in a more subjective way. In typical formulations a smaller value of the outcome (delay) means a better effect (higher service quality or client satisfaction). Otherwise, the outcomes can be replaced with their complements to some large number. Therefore, without loss of generality, we can assume that each individual outcome y_i is to be minimized which results in a multiple criteria minimization model.

The simplest services structure forms the uniform problem where each service represents a single unit. Usually, one is interested in putting into allocation model some additional demand weights $w_i > 0$ to represent the amount of demand for the specific service. Integer weights can be interpreted as numbers of unweighted identical services to be repeated independently. For initial theoretical considerations we will assume that the problem is transformed (disaggregated) to the uniform one (that means all the demand weights are equal to 1). Note that such a disaggregation is possible for integer as well as rational demand weights, but it usually dramatically increases the problem size. Therefore, we consider solution concepts which can be applied directly to the weighted problem. For this purpose we will use the normalized demand weights

$$\bar{w}_i = w_i / \sum_{j=1}^{m} w_j \quad \text{for } i = 1,2,\dots,m \tag{1}$$

rather than the original quantities w_i. Note that, in the case of uniform problem (all $w_i = 1$), all the normalized weights are given as $\bar{w}_i = 1/m$.

Telecommunication problems deal with routing of the data traffic in an existing network or with designing the network expansions to accommodate the traffic. Both type of problems require the allocation of network resources (capacities or potential capacities). Let us consider a connected network consisted of a node set N to represent various locations. Directed links $(j,k) \in L \subset N \times N$ are attributed by the bandwidth/capacity coefficients b_{jk} and the delay/distance/cost coefficients c_{jk}. Further, we consider a set $I = \{1,2,\dots,m\}$ of m services. Each service is related to some data traffic between two network nodes. Thus, the service is described by a directed pair of nodes $\big(s(i), d(i)\big)$ representing the source and the destination of the data traffic, respectively. The amount of the data traffic related to service i is described by the demand weight w_i. The latter may be skipped in the case of uniform problem where all the services generate the same amount of data traffic ($w_i = 1$ for all i).

Within a telecommunication network the data traffic is generated by a huge number of nodes exchanging data. In such a network, a relatively small subset $H \subset N$ of nodes are chosen to serve as hubs which can be used as intermediate switching points [1, 6]. Given a set of hubs, data traffic generated by a service is sent from the source node to a hub first. It can be then sent along communications link between hubs, and finally reach the destination node along a link from a hub. The hub-based network organization allows the data traffic to be consolidated on the inter-hub links.

While taking into account the hub-based network structure, the main decisions to be made for the services organization can be described with the assignment of a directed pair of hubs $\big(h'(i), h''(i)\big)$ to each service i. The data traffic for service i is then implemented by sending from the source $s(i)$ to the hub $h'(i)$ first, the use of the inter-hub connection from $h'(i)$ to $h''(i)$ next, and the final sending from $h''(i)$ to the service destination $d(i)$. The delay/distance of such a data path is usually assumed to be defined as the sum of several link delays $c_{s(i),h'(i)} + c_{h'(i),h''(i)} + c_{h''(i),d(i)}$. Note that a single hub can be used is some cases $\big(h'(i) = h''(i)\big)$ which may require a definition of the corresponding dummy inter-hub links.

In the case of the demand weights for various services there is no justification for a strict assignment of a single path to the specific service since several units may be sent along different paths. Therefore, the main decisions may be modeled with variables $x_{ijk}\big(i \in I;\ j,k \in H\big)$ expressing the amount of data traffic related to service i routed via hubs h_j and h_k. To meet the problem requirements, the decision variables x_{ijk} have to satisfy the following constraints:

$$\sum_{j \in H} \sum_{k \in H} x_{ijk} = w_i \qquad \text{for } i \in I, \tag{2}$$

$$\sum_{i \in I} x_{ijk} \le b_{jk} \qquad \text{for } j,k \in H, \tag{3}$$

$$x_{ijk} \ge 0 \qquad \text{for } i \in I;\ j,k \in H, \tag{4}$$

where Eqs. (2) guarantee the routing of whole service demands while inequalities (3) keep the data traffic within the capacity limits. Note that taking into account the hub-based network specificity we have considered the capacity constraints only for the inter-hub links.

The unit performance measure (delay) of the service i may be expressed with the following linear function:

$$f_i(\mathbf{x}) = \frac{1}{w_i} \sum_{j,k \in H} \big[c_{s(i),j} + c_{jk} + c_{k,d(i)}\big] x_{ijk} \quad \text{for} \quad i \in I. \tag{5}$$

Hence, all the functions $f_i(\mathbf{x})$ need to be minimized. The typical problems involving routing decisions are considered as dynamic and stochastic. Nevertheless, one may analyze a straightforward static allocation problem related to traffic engineering (routing) decisions within a telecommunication (or transportation) network. Such a problem depends, simply, on minimization of criteria (5) subject to constraints (2)–(4).

Similar model may be considered for the inter-hub bandwidth allocation problem related to the network design issues. Namely, one may minimize the same criteria (5) and the same constraints (2)–(4), but the bandwidth b_{jk} in constraints (3) need to be considered decision variables rather than data parameters. Again, it is a straightforward network design model but its analysis may be useful at some initial phases of the design process.

In the above model we allow the services to be partitioned in various portions of the demand and implemented with possibly different routing. We believe that is acceptable for most applications related to data transfer as the standard data package is relatively extremely small when comparing to the total amount of demand. Moreover, new routing protocols developed for the Internet services, like the multi-protocol label switching, allow much flexibility in the traffic engineering solutions [4]. Nevertheless, the problem (2)–(5) may be adapted to the requirement of a single route assigned to each service, if necessary. For this purpose, one needs to introduce binary decision variables x_{ijk} equal 1 when the inter-hub link (h_j,h_k) is used to implement service i, and 0 otherwise. The constraints take then the following form:

$$\sum_{j\in H}\sum_{k\in H} x_{ijk} = 1 \qquad \text{for} \quad i\in I,$$

$$\sum_{i\in I} w_i x_{ijk} \le b_{jk} \qquad \text{for} \quad j,k\in H$$

and the resulting model is very close to the location problems [6, 10].

In problem (2)–(5) we have considered all the hubs as directly connected by the corresponding inter-hub links. In telecommunication networks hubs are rather organized in some network structure (architecture) which causes the existence of some interactions (common bandwidth limits) between various inter-hub connections representing rather paths (routes) than direct links. The modern telecommunication networks heavily use the architecture of a collection of bidirectional rings (as in SONET standard [3]). Below we specify in details such an allocation model where the hubs are arranged in a cycle and the traffic engineering problem needs to take into account the bidirectional ring-loading issues. This type of models we will use in Section 5 to demonstrate some computational results.

Let us consider again a connected network consisted of a node set N directed links $(j,k)\in L\subset N\times N$ which are attributed by the bandwidth/capacity coefficients b_{jk} and the delay/distance/cost coefficients c_{jk}. A set $I=\{1,2,\dots,m\}$ of m services is considered. Each service is related to a directed pair of nodes $(s(i),d(i))$ (the source and the destination of the data traffic), and it requires the amount w_i of the data traffic (demand weight w_i). A relatively small subset $H\subset N$ of p nodes are chosen to serve as hubs. Hubs h_1, h_2, …, h_{p-1}, h_p are arranged clockwise in a cycle (ring). That means, there are p clockwise directed inter-hub links: (h_1,h_2),(h_2,h_3), …, (h_{p-1},h_p), (h_p,h_1), and p counterclockwise directed inter-hub links: (h_p,h_{p-1}),(h_{p-1},h_{p-2}), …, (h_2,h_1), (h_1,h_p).

The main decisions may be modeled with variables x'_{ijk} and x''_{ijk} $(i\in I;\ j,k\in H)$ expressing the amount of data traffic related to service i routed via hubs h_j and h_k using clockwise or counterclockwise connection, respectively. To meet the service demand requirements, the decision variables have to satisfy the following constraints:

$$\sum_{j\in H}\sum_{k\in H}(x'_{ijk}+x''_{ijk}) = w_i \qquad \text{for } i\in I, \tag{6}$$

$$x'_{ijk}, x''_{ijk} \ge 0 \qquad \text{for } i\in I;\ j,k\in H. \tag{7}$$

Recall that there is a piece of data traffic which passes trough a single hub not generating the ring traffic either clockwise or counterclockwise. Namely, $x'_{ijj}+x''_{ijj}$ for $j\in H$ is the amount of such traffic and it could be represented by a single variable but we have accepted the redundancy to keep the constraints (6) simpler.

To analyze the bandwidth (links capacity) allocation one needs to accumulate the traffic load of specific links in the ring. Let $(l_1,l_2)\in C$ denote a clockwise link in the ring, i.e. $l_2=l_1+1$ for $l_1=1,\dots,p-1$ or $l_2=1$ for $l_1=p$. The link is loaded with clockwise traffic from hub h_j to hub h_k for $j=1,\dots,l_2-1$ and $k=l_2,\dots,p$ or $k=1,\dots,j-1$ as well as (if $l_2\le p-1$) for $j=l_2+1,\dots,p$ and $k=l_2,\dots,j-1$. Hence, the clockwise traffic of all the services generates the following (clockwise) link load

$$z'_{l_1,l_2} = \sum_{i\in I}\left[\sum_{j=1}^{l_2-1}\Big(\sum_{k=l_2}^{p} x'_{ijk} + \sum_{k=1}^{j-1} x'_{ijk}\Big) + \sum_{j=l_2+1}^{p}\sum_{k=l_2}^{j-1} x'_{ijk}\right] \tag{8}$$

for each $(l_1,l_2)\in C$. By symmetry, the counterclockwise traffic of all the services generates the (counterclockwise) link load

$$z''_{l_1,l_2} = \sum_{i\in I}\left[\sum_{k=1}^{l_2-1}\Big(\sum_{j=l_2}^{p} x''_{ijk} + \sum_{j=1}^{k-1} x''_{ijk}\Big) + \sum_{k=l_2+1}^{p}\sum_{j=l_2}^{k-1} x''_{ijk}\right] \tag{9}$$

for each $(l_1,l_2)\in C$. Note that z''_{l_1,l_2} denotes, in fact, the load of directed counterclockwise link (l_2,l_1). With commonly considered bidirectional capacity (bandwidth) limits the link loads must satisfy the constraints

$$z'_{l_1,l_2} + z''_{l_1,l_2} \le b_{l_1,l_2} \text{for } (l_1,l_2)\in C. \tag{10}$$

In the case of independently considered separate single-directional capacity limits, the latter needs to be replaced with constraints

$$z'_{l_1,l_2} \le b'_{l_1,l_2} \quad \text{and} \quad z''_{l_1,l_2} \le b''_{l_1,l_2} \quad \text{for} \quad (l_1,l_2)\in C.$$

The unit performance measure (delay) for the service i is expressed with the following linear function:

$$f_i(\mathbf{x}) = \frac{1}{w_i}\sum_{j,k\in H}(c_{s(i),j} + d'_{jk} + c_{k,d(i)})x'_{ijk} + \\ + \frac{1}{w_i}\sum_{j,k\in H}(c_{s(i),j} + d''_{jk} + c_{k,d(i)})x''_{ijk}\,, \qquad (11)$$

where d'_{jk} and d''_{jk} denote the delays along the clockwise and counterclockwise, respectively, paths from h_j to h_k in the ring C. For instance, in the case of $1 \le j < k \le p$ one gets $d'_{jk} = c_{j,j+1} + \ldots + c_{k-1,k}$. Certainly, all the functions $f_i(\mathbf{x})$ need to be minimized. Hence, a simple traffic engineering problem with bidirectional ring loading issues can be considered as multiple criteria minimization of (11) subject to constraints (6)–(10).

The problem (6)–(11) may be adapted to the requirement of a single inter-hub route assigned to each service, if necessary. Let us assume that the data traffic related to service i and routed via hubs h_j and h_k has to use either clockwise or counterclockwise connection without any splitting. This requirement can be modeled by introducing binary decision variables r_{ijk} equal 1 when the clockwise connection from h_j to h_k is used to implement service i, and 0 for the counterclockwise connection. The model needs to be extended then with the constraints of the following form:

$$x'_{ijk} \le w_i r_{ijk} \quad \text{and} \quad x'_{ijk} \le w_i(1 - r_{ijk}) \quad \text{for} \quad i \in I,\ j,k \in H\,.$$

One may also formulate a network design problem where quantities b_{l_1,l_2} for $(l_1,l_2) \in C$ are considered as a set of multiple criteria to be minimized subject to constraints (6)–(10) with possible upper limits on service delays $f_i(\mathbf{x})$.

3. The solution concept

Assuming that the generic allocation problem has been disaggregated to the unweighted form (all $w_i = 1$), it may be stated as the following multiple criteria minimization problem:

$$\min\ \{\mathbf{f}(\mathbf{x}) : \mathbf{x} \in Q\}\,, \qquad (12)$$

where $\mathbf{f} = (f_1,\ldots,f_m)$ is a vector of the individual objective functions which measure the outcome (effect) $y_i = f_i(\mathbf{x})$ of the allocation pattern $\mathbf{x}$ for service i.

We do not assume any special form of the feasible set while introducing the solution concepts. We rather allow the feasible set to be a general, possibly discrete (nonconvex), set. Similarly, we do not assume any special form of the individual objective functions nor their special properties (like convexity). Therefore, the solution concepts may be applied to various allocation problems. Nevertheless, the solution concepts, we consider, are implementable by a linear objective and a number of auxiliary linear inequalities. Thus the solution concepts preserve a possible structure (LP or convexity) of the allocation problem under analysis.

Most classical allocation studies focus on the minimization of the mean (or total) outcome or the minimization of the maximum (the worst) outcome. Both the corresponding solution concepts are well defined for aggregated allocation models using demand weights $w_i > 0$. Exactly, for the weighted allocation problem, the *minisum* solution concept is defined by the minimization of the objective function expressing the *mean* (average) outcome

$$\mu(\mathbf{y}) = \sum_{i=1}^{m} \bar{w}_i y_i$$

but it is also equivalent to the minimization of the total outcome $\sum_{i=1}^{m} w_i y_i$. The *minimax* solution concept is defined by the minimization of the objective function representing the *maximum* (worst) outcome

$$M(\mathbf{y}) = \max_{i=1,\ldots,m} y_i$$

and it is not affected by the demand weights at all. Both the classical solution concepts are represented with simple aggregation of multiple criteria model (12). Namely, the minisum approach simply use the weighted sum of criteria

$$\min\Big\{\sum_{i=1}^{m} \bar{w}_i f_i(\mathbf{x}) : \mathbf{x} \in Q\Big\} \qquad (13)$$

while the minimax approach results in a problem

$$\min\big\{t : \mathbf{x} \in Q;\ t \ge f_i(\mathbf{x}) \quad \text{for}\ i = 1,2,\ldots,m\big\} \qquad (14)$$

with only one auxiliary variable t and m inequalities to define it.

Since the minisum approach is based on averaging, it often provides solutions where low demand services related to remote destinations are discriminated in terms of delays. On the other hand, allocating the resources to optimize the worst case may cause a large increase in the total delays thus generating a substantial loss in the overall system efficiency. This has led to a search for some compromise solution concept.

A natural generalization of the maximum (worst) outcome $M(\mathbf{y})$ is the worst conditional mean defined as the mean within the specified tolerance level (amount) of the worst outcomes. For the simplest case of the unweighted allocation problem (12), one may simply define the worst conditional mean $M_{\frac{k}{m}}(\mathbf{y})$ as the mean outcome for the k worst-off services (or rather k/m portion of the worst services). This can be mathematically formalized as follows. First, we introduce the ordering map $\Theta : R^m \to R^m$ such that $\Theta(\mathbf{y}) = (\theta_1(\mathbf{y}), \theta_2(\mathbf{y}), \ldots, \theta_m(\mathbf{y}))$, where $\theta_1(\mathbf{y}) \ge \theta_2(\mathbf{y}) \ge \cdots \ge \theta_m(\mathbf{y})$ and there exists a permutation τ of set I such that $\theta_i(\mathbf{y}) = y_{\tau(i)}$ for $i = 1,2,\ldots,m$. The use of ordered outcome vectors $\Theta(\mathbf{y})$ allows us to focus on distributions of outcomes impartially. Next, the linear cumulative map is applied to ordered outcome vectors to get $\bar{\Theta}(\mathbf{y}) = (\bar{\theta}_1(\mathbf{y}), \bar{\theta}_2(\mathbf{y}), \ldots, \bar{\theta}_m(\mathbf{y}))$ defined as

$$\bar{\theta}_k(\mathbf{y}) = \sum_{i=1}^{k} \theta_i(\mathbf{y}), \quad \text{for}\ k = 1,2,\ldots,m. \qquad (15)$$

The coefficients of vector $\overline{\Theta}(\mathbf{y})$ express, respectively: the largest outcome, the total of the two largest outcomes, the total of the three largest outcomes, etc. Hence, the *worst k/m–conditional mean* $M_{\frac{k}{m}}(\mathbf{y})$ is given as

$$M_{\frac{k}{m}}(\mathbf{y}) = \frac{1}{k}\bar{\theta}_k(\mathbf{y}), \quad \text{for } k = 1,2,\dots,m. \qquad (16)$$

Note that for $k = 1$, $M_{\frac{1}{m}}(\mathbf{y}) = \bar{\theta}_1(\mathbf{y}) = \theta_1(\mathbf{y}) = M(\mathbf{y})$ thus representing the maximum outcome, and for $k = m$, $M_1(\mathbf{y}) = \frac{1}{m}\bar{\theta}_m(\mathbf{y}) = \frac{1}{m}\sum_{i=1}^m \theta_i(\mathbf{y}) = \frac{1}{m}\sum_{i=1}^m y_i = \mu(\mathbf{y})$ which is the mean outcome. Except for these two limiting cases, the definition (16) is hardly implementable due to the use of the ordering operator. The following theorem shows that the worst conditional mean can be found by minimization of a scalar piecewise linear convex function.

Theorem 1. For any vector $\mathbf{y} \in R^m$ the corresponding quantity $\bar{\theta}_k(\mathbf{y})$ represents the minimum value of the (scalar) optimization:

$$\bar{\theta}_k(\mathbf{y}) = \min_{t\in R} \tfrac{1}{m}\sum_{i=1}^m \Big[k(t-y_i)_+ + $$
$$+(m-k)(y_i-t)_+\Big] + \tfrac{k}{m}\sum_{i=1}^m y_i \qquad (17)$$

while $\bar{t} = \theta_k(\mathbf{y})$ is an optimal solution (argument) of the above optimization.

Proof. First, we show that $\bar{t} = \theta_k(\mathbf{y})$ minimizes the function:

$$g_k(t) = \sum_{i=1}^m \Big[k(t-y_i)_+ + (m-k)(y_i-t)_+\Big]. \qquad (18)$$

Note that $g_k(t) = \sum_{i=1}^m \Big[k(t-\theta_i(\mathbf{y}))_+ + (m-k)(\theta_i(\mathbf{y})-t)_+\Big]$. Consider $t = \bar{t}+\delta$ with any $\delta \in R$ (positive or negative). For $i = k+1,\dots,m$

$$\big((\bar{t}+\delta)-\theta_i(\mathbf{y})\big)_+ \ge \big(\bar{t}-\theta_i(\mathbf{y})\big)_+ + \delta$$

and

$$\big(\theta_i(\mathbf{y})-(\bar{t}+\delta)\big)_+ \ge 0\ ,$$

while for $i = 1,\dots,k$

$$(\theta_i(\mathbf{y})-(\bar{t}+\delta))_+ \ge (\theta_i(\mathbf{y})-\bar{t})_+ - \delta$$

and

$$\big((\bar{t}+\delta)-\theta_i(\mathbf{y})\big)_+ \ge 0\ .$$

Hence, one gets

$$k\sum_{i=1}^m \big((\bar{t}+\delta)-y_i\big)_+ \ge k\sum_{i=1}^m (\bar{t}-y_i)_+ + k(m-k)\delta$$

and

$$(m-k)\sum_{i=1}^m \big(y_i-(\bar{t}+\delta)\big)_+ \ge (m-k)\sum_{i=1}^m (y_i-\bar{t})_+ +$$
$$-\ (m-k)k\delta.$$

Thus finally, $g_k(\bar{t}) \le g_k(\bar{t}+\delta)$ for all $\delta \in R$.

Further, calculating the minimal value of (18), we get:

$$g_k\big(\theta_k(\mathbf{y})\big) =$$
$$= \sum_{i=1}^m \Big[k\big(\theta_k(\mathbf{y})-y_i\big)_+ + \big(m-k\big)\big(y_i-\theta_k(\mathbf{y})\big)_+\Big] =$$
$$= k\sum_{i=k+1}^m \big(\theta_k(\mathbf{y})-\theta_i(\mathbf{y})\big) - (m-k)\sum_{i=1}^k \big(\theta_k(\mathbf{y})-\theta_i(\mathbf{y})\big) =$$
$$= k\sum_{i=1}^m \big(\theta_k(\mathbf{y})-\theta_i(\mathbf{y})\big) - m\sum_{i=1}^k \big(\theta_k(\mathbf{y})-\theta_i(\mathbf{y})\big) =$$
$$= km\theta_k(\mathbf{y}) - k\sum_{i=1}^m \theta_i(\mathbf{y}) - mk\theta_k(\mathbf{y}) + m\sum_{i=1}^k \theta_i(\mathbf{y}) =$$
$$= m\sum_{i=1}^k \theta_i(\mathbf{y}) - k\sum_{i=1}^m \theta_i(\mathbf{y}) = m\bar{\theta}_k(\mathbf{y}) - k\sum_{i=1}^m y_i\,.$$

Hence

$$\bar{\theta}_k(\mathbf{y}) = \frac{1}{m}g_k\big(\theta_k(\mathbf{y})\big) + \frac{k}{m}\sum_{i=1}^m y_i = \frac{1}{m}\min_{t\in R} g_k(t) + \frac{k}{m}\sum_{i=1}^m y_i$$

which completes the proof of (17).

It follows from Theorem 1 that, for a given vector $\mathbf{y}$, the value of $\bar{\theta}_k(\mathbf{y})$ may be found by solving the linear program:

$$\bar{\theta}_k(\mathbf{y}) = \min\ \sum_{i=1}^m \Big(\frac{k}{m}d_i^- + \frac{m-k}{m}d_i^+\Big) + \frac{k}{m}\sum_{i=1}^m y_i$$
$$\text{subject to}$$
$$d_i^+ - d_i^- = y_i - t, \quad d_i^+, d_i^- \ge 0\ \forall i,$$

where t is an unbounded variable representing a freely selected target while nonnegative variables d_i^+ and d_i^- represent, for several outcome values y_i, their upside and downside deviations from the selected target t, respectively. Moreover, the target variable t takes the value of $\theta_k(\mathbf{y})$ at the optimal solution. The linear program can be further simplified by the elimination of variables d_i^- representing the downside deviations. Hence, following (16), the worst k/m–conditional mean $M_{\frac{k}{m}}(\mathbf{y})$, for $k = 1,2,\dots,m$, is given by the following optimization:

$$M_{\frac{k}{m}}(\mathbf{y}) = \min\Big\{t + \frac{1}{k}\sum_{i=1}^m d_i^+ : d_i^+ \ge y_i - t,\ d_i^+ \ge 0 \forall i\Big\}. \qquad (19)$$

This allows us to define *the k/m–conditional minimax* solution for the unweighted allocation problem (12) as the optimal solution to the optimization problem:

$$\min\Big\{t + \frac{1}{k}\sum_{i=1}^m d_i^+ : \mathbf{x} \in Q; \quad d_i^+ \ge f_i(\mathbf{x}) - t,\ d_i^+ \ge 0\ \forall i\Big\} \qquad (20)$$

or simply

$$\min\Big\{t + \frac{1}{k}\sum_{i=1}^m \big(f_i(\mathbf{x})-t\big)^+ : \mathbf{x} \in Q\Big\}\,,$$

where $(.)^+$ denotes the nonnegative part of a number.

One may notice that formula (17) in Theorem 1 as well as the subsequent optimization problems (19) or (20) defining the conditional minimax, all they are given directly on outcomes y_i without any use of the ordering operator Θ. Thus, in the case of a weighted allocation problem, Theorem 1 applied to the corresponding disaggregated problem (with equal weights) results in formulas allowing us to reaggregate the outcomes related to the same services. Hence, in the presence of demand weights $w_i > 0$, for any real tolerance level $0 < \beta \leq 1$, there is well defined the *worst β–conditional mean*

$$M_\beta(\mathbf{y}) = \min\Big\{t + \frac{1}{\beta}\sum_{i=1}^{m} \bar{w}_i d_i^+ : d_i^+ \geq y_i - t,\ d_i^+ \geq 0\ \forall i\Big\},$$

where $\bar{w}_i$ denote the normalized weights (1). This allows us to define the *β–conditional minimax* solution for the weighted allocation problem as the optimal solution to the following problem:

$$\min\Big\{t + \frac{1}{\beta}\sum_{i=1}^{m} \bar{w}_i d_i^+ : \mathbf{x} \in Q;\ d_i^+ \geq f_i(\mathbf{x}) - t,\ d_i^+ \geq 0\ \forall i\Big\}. \tag{21}$$

Note that (21) uses $m+1$ auxiliary variables and m inequalities to define minimization of the worst conditional mean. When the tolerance level β tends to 0, then all the deviational variables d_i^+ are forced to 0. Therefore, the limiting problem of the standard minimax optimization takes the simpler form (14). On the other hand, for $\beta = 1$, problem (21) takes the form

$$\min\Big\{\sum_{i=1}^{m} \bar{w}_i(d_i^+ + t) : \mathbf{x} \in Q;\ d_i^+ + t \geq f_i(\mathbf{x}),\ d_i^+ \geq 0\ \forall i\Big\},$$

which can be simplified to the standard minisum optimization (13).

The cumulative ordered outcomes (15), used to introduce the worst conditional mean, are closely related with the Pigou-Dalton theory of inequality measurement [12] and the Lorenz curves. Assume that the allocation problem (12) is transformed (disaggregated) into the unweighted one (that means all the demand weights are equal to 1). Vector $\bar{\Theta}(\mathbf{y})$ $\big($exactly $\frac{1}{m}\bar{\Theta}(\mathbf{y})\big)$ can be viewed graphically with the curve connecting point (0,0) and points $(i/m, \bar{\theta}_i(\mathbf{y})/m)$ for $i = 1, 2, \ldots, m$. Graphs of vectors $\bar{\Theta}(\mathbf{y})$ take the form of unnormalized concave curves, the *(upper) absolute Lorenz curves*.

The absolute Lorenz curves defines the relation (partial order) of the equitable dominance. The equitable dominance is originally defined by axioms of efficiency, impartiality and the Pigou-Dalton principle of transfers [7, 10]. Nevertheless, due to the results of the majorization theory [7], it can be expressed with inequalities on the absolute Lorenz curves. Exactly, outcome vector $\mathbf{y}' \in Y$ equitably dominates $\mathbf{y}'' \in Y$, if and only if $\bar{\theta}_i(\mathbf{y}') \leq \bar{\theta}_i(\mathbf{y}'')$ for all $i \in I$ where at least one strict inequality holds. We say that an allocation pattern $\mathbf{x} \in Q$ is *equitably efficient* (is an equitably efficient solution of the multiple criteria problem (12)), if and only if there does not exist any $\mathbf{x}' \in Q$ such that $\mathbf{f}(x')$ equitably dominates $\mathbf{f}(\mathbf{x})$. Note that with the relation of equitable dominance an outcome vector of small unequal outcomes may be preferred to an outcome vector with large equal outcomes. Each equitably efficient solution is also an efficient solution but not *vice verse*.

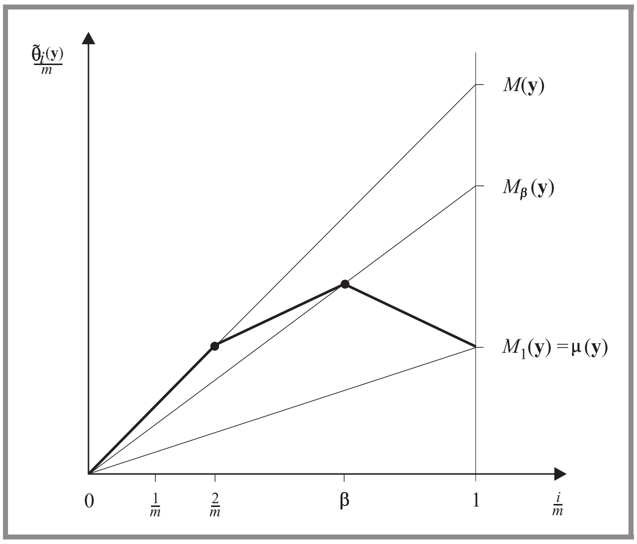


Fig. 1. Absolute Lorenz curve and the worst conditional means.

Recall that the worst conditional mean is defined as $M_{\frac{k}{m}}(\mathbf{y}) = \bar{\theta}_k(\mathbf{y})/k$ while vector $\frac{1}{m}\bar{\Theta}(\mathbf{y})$ can be viewed graphically with the upper absolute Lorenz curve connecting point (0,0) and points $(i/m, \bar{\theta}_i(\mathbf{y})/m)$ for $i = 1, 2, \ldots, m$. Hence, as shown in Fig. 1, the worst conditional mean represents the projection of the point of the Lorenz curve onto the vertical line at point 1 ($i = m$). This also demonstrates that for any given outcome vector $\mathbf{y}$, the worst conditional mean $M_\beta(\mathbf{y})$ is monotonic (nonincreasing), when considered as a function of β, i.e. $0 < \beta' \leq \beta'' \leq 1$ implies $M_{\beta'}(\mathbf{y}) \geq M_{\beta''}(\mathbf{y})$. Further, since the worst conditional mean $M_\beta(\mathbf{y})$ is a quantity proportional to the value of the absolute Lorenz curve at a specific point β, comparison of the worst conditional means (for the same given β) is consistent with the equitable dominance. Exactly, this leads to the following assertion.

Theorem 2. Except for allocation patterns with identical the worst conditional means $M_\beta(\mathbf{y})$, every allocation pattern $\mathbf{x} \in Q$ that is minimal for $M_\beta\big(\mathbf{f}(\mathbf{x})\big)$ is an equitably efficient solution of the allocation problem (12).

4. Computational results

In this section we report some results of our initial computational experience with the conditional minimax solution concept applied to traffic engineering problems. We have solved randomly generated problems following the formulation from Section 2. Thus, our analysis is limited to a simple allocation model where the hubs are arranged in a ring and the traffic engineering problem needs to take into account the bidirectional ring-loading issues.

Our computational tests are based on the randomly generated problems (6)–(11). The generation procedure works as follows. First, a ring with a given number of hubs is built. The clockwise and counterclockwise inter-hub links are distinguished. The delays for these links are generated as random integers uniformly distributed between 5 and 10. Having the ring defined, a given number of services is randomly generated. For each service i, the source node $s(i)$ as well as the destination node $d(i)$ are linked to uniquely selected hub each. The pair of hubs for the given service is chosen randomly from all hubs in the ring, excluding the case of the source node and the destination node attached to the same hub. The delays of links between the source or the destination nodes and their respective hubs in the ring are randomly generated as integers uniformly distributed between 10 and 20. Finally, the demands w_i for the services are generated as random integers uniformly distributed between 1 and 100. All the inter-hub links are assumed to have the same bandwidth. The bandwidth value is defined as a result of the following procedure. We start with initial bandwidth defined as $\sum_{i=1}^{m} w_i$ to guarantee the feasibility (solvability) of the generated problem. Further, we try to reduce the bandwidth still preserving the feasibility. For this purpose, 8 steps of the bisection procedure is applied whereas the current bandwidth is decreased or increased depending on the feasibility of the problem. This allows us to built nontrivial feasible bidirectional ring-loading problems.

The solution concept of conditional minimax provides a compromise between the minimax and the minisum approaches. Table 1 shows the quality of this compromise. It provides average percentage distribution of delays for conditional minimax solutions obtained by varying tolerance level β in the objective M_β. Distribution is calculated as an average of 100 randomly generated problems with 20 hubs and 8 services. Resulting delays are partitioned into clusters of range ten: $[20;30)$, $[30;40)$ etc. Each row represents average distribution for a particular tolerance level β. Exactly, each field gives the percentage of delays within a given range in 100 optimal solutions. It is clear that percentage of low delays increases with β (left columns). On the other hand, for small values of the tolerance level β, the percentage of large delays is forced to zero. With β increasing, large delays begin to occur, first incidentally like delays over 120 for $\beta = 0.1$ or 0.2 (resulting in average percentage below 0.1%), next with a raising percentage.

Table 1
Average distributions of delays for 100 random problems

β	Average percentage of delays for β–conditional minimax solutions												
	20	30	40	50	60	70	80	90	100	110	120	130	140
0.1	2.3	9.0	9.3	8.3	10.5	12.2	12.8	15.9	16.3	3.4	0.0		
0.2	2.1	8.9	8.8	8.4	10.0	10.9	16.3	15.8	15.8	2.9	0.0		
0.3	2.3	8.6	7.9	7.6	11.5	14.5	15.3	16.8	11.7	3.3	0.3	0.1	
0.4	2.3	8.6	8.4	7.9	11.8	15.4	14.3	16.6	10.7	3.5	0.5	0.1	
0.5	2.3	8.6	9.0	8.3	12.8	15.7	13.9	14.2	10.6	4.1	0.5	0.1	
0.6	2.3	8.6	10.7	8.9	12.3	15.4	12.9	13.1	9.9	5.1	0.6	0.2	
0.7	2.3	9.2	11.3	11.3	12.3	14.2	10.3	11.1	11.2	5.2	1.1	0.2	0.2
0.8	2.3	10.1	11.9	12.2	12.9	11.9	8.7	11.3	11.2	5.7	1.3	0.2	0.2
0.9	2.3	10.6	13.4	11.0	11.9	12.2	8.3	10.9	11.9	5.8	1.1	0.4	0.2
1.0	2.3	10.8	13.6	10.8	11.6	12.2	8.3	10.4	12.3	5.9	0.9	0.6	0.2

The main properties of the conditional minimax solution concepts are visible in averages for 100 problems. Nevertheless, a single problem allows us to demonstrate much better the differences among the solutions. Therefore, we have selected and analyzed in details one of the randomly generated problems. Table 2 shows the resulting distributions of delays for four various conditional minimax solution concepts applied to this sample problem. One may notice that the distributions of delays are significantly different despite their means are quite close.

Table 2
Distributions of delays for a sample problem

β	Percentage distribution of delays								μ	M
	30	40	50	60	70	80	100	120		
0.1	22.3	0.2	5.1	22.6	16.7	33.0			65.64	85.35
0.4	22.3	0.2	5.1		50.9	21.4			66.00	85.35
0.7	22.3	0.2	39.3		16.7		21.4		65.43	101.74
1.0	22.3	34.4	5.1		16.7			21.4	64.46	124.86

The distributions of delays generated by several solutions from Table 2 are also presented graphically. In Fig. 2, for each of four distributions of delays the values of all the worst conditional means are shown as functions of the tolerance level. This results in four curves, each starting from the corresponding maximum delay and reaching the mean delay when the tolerance level tends to 1. One may notice that among our four solutions the 0.4-conditional minimax has the smallest worst conditional mean for tolerance levels between 0.21 and 0.6 as well as it remains an alternative optimal solution to the minimax solution for smaller tolerance levels.

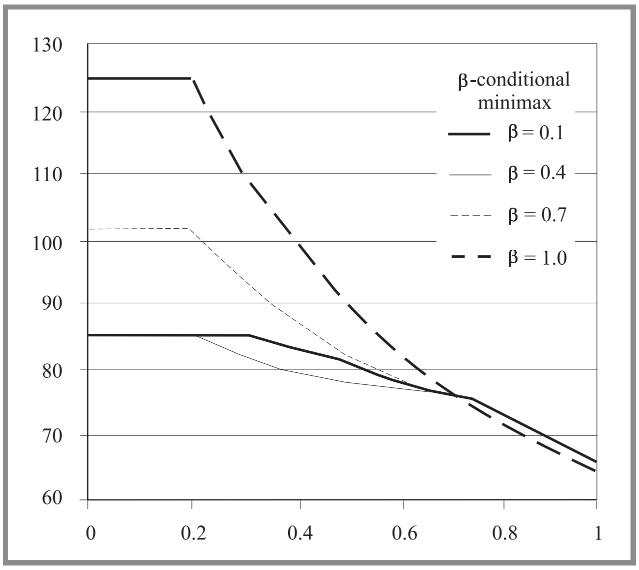


Fig. 2. Curves of the worst conditional means.

Figure 3 shows the absolute Lorenz curves built for the distributions of delays for our four conditional minimax solutions of the sample problem. One might notice from Table 2 that the 0.4–conditional minimax generates the same maximum delay as the 0.1-conditional minimax while its mean is greater than that of the latter. Hence, while dealing with only two criteria of the maximum delay and the mean delay, the 0.4-conditional minimax solution is dominated.

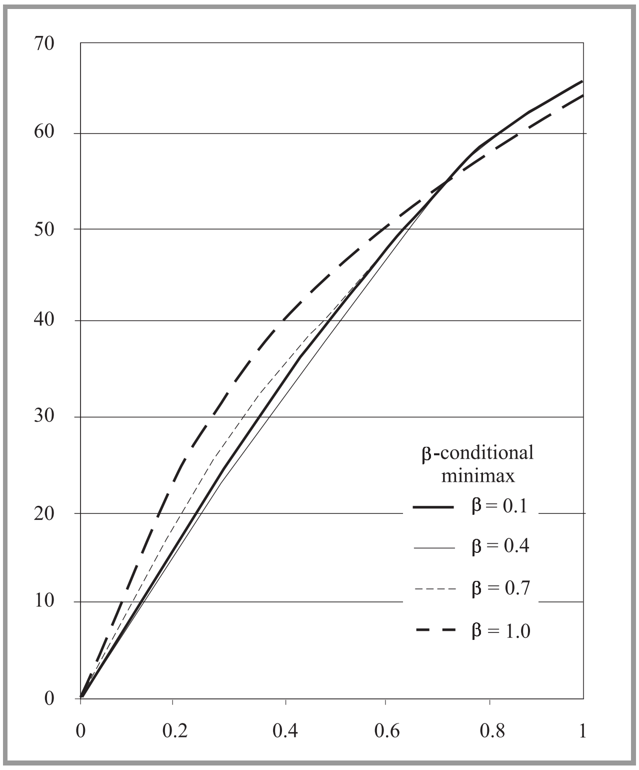


Fig. 3. Absolute Lorenz curves for the sample problem.

Nevertheless, as shown with the absolute Lorenz curves, it is equitably nondominated and the minimization of the worst conditional mean with the tolerance level β between 0.21 and 0.6 points out this solution as optimal.

Table 3
Average solution times for the 0.5-conditional minimax

Hubs	Number of services (m)			
p	50	100	200	500
50	0.10	0.40	0.60	3.20
100	0.40	0.60	1.40	5.40
200	0.40	1.20	2.40	11.00
500	1.20	3.40	6.40	38.60

We tested solution times for different number of services m and number of hubs p. For each specified size parameters we generated randomly 5 problems (6)–(11). The 0.5-conditional minimax solutions were then found. All computations were performed on a PC with the Pentium 200 MHz processor employing the CPLEX 6.0 package [5]. The results are presented in Table 3. Every reported time is an average of 5 results (in seconds) for problems of the given size. One may notice that even problems with 500 hubs were solved very fast.

5. Concluding remarks

Resource allocation problems are concerned with the allocation of limited resources among competing services or other activities so as to achieve the best overall performances. In various systems which serve many users, like in telecommunication systems, there is a need to allocate resources equitably among the competing services. In this paper we have developed an equitable solution concept of the conditional minimax. Although similar to the standard minimax approach, the conditional minimax takes into account the amount of services related to the worst performances. For a specified tolerance level (portion of services amount) β we take into account the entire group of the β portion maximum results and we consider their average as the worst conditional mean to be minimized. According to this definition the solution concept is based on averaging restricted to the group of the worst performances defined by the tolerance level. Hence, by the selection of the tolerance level various equitable preferences may be modeled.

The solution concept of the conditional minimax, similar to the standard minimax approach, can be defined by optimization of a linear objective and a number of auxiliary linear inequalities. Therefore, the concept may be effectively applied to various resource allocation problems. Our initial computational experiments with the conditional minimax applied to a straightforward traffic engineering model (restricted to a single ring bidirectional loading) confirm the theoretical properties of the solution concept. Bidirectional ring loading problems containing up 500 hubs were solved very fast with the general purpose LP solver. Nevertheless, many specific large-scale allocation models (especially discrete ones) may need some specialized exact or approximate algorithms. Thus, further research on computational aspects of the conditional minimax solution concept is necessary.

References

[1] J. F. Campbell, "Hub location and the *p*-hub median problem", *Oper. Res.*, vol. 44, pp. 923–935, 1996.

[2] D. W. Corne, M. J. Oates, and G. D. Smith, Eds., *Telecommunications Optimization: Heuristics and Adaptive Techniques.* New York: Wiley, 2000.

[3] S. Cosares, D. N. Deutsch, I. Saniee, and O. J. Wasem, "SONET toolkit: a decision support system for designing robust and cost-effective fiber-optic networks", *Interfaces*, vol. 25, pp. 20–40, 1995.

[4] E. Gourdin, "Optimizing Internet networks", *OR/MS Today*, vol. 28, pp. 46–49, 2001.

[5] ILOG Inc., Using the CPLEX Callable Library. Incline Village: ILOG Inc., CPLEX Division, 1997.

[6] J. G. Klincewicz, "Hub location in backbone/tributary network design: a review", *Loc. Sci.*, vol. 6, pp. 307–335, 1998.

[7] M. M. Kostreva and W. Ogryczak, "Linear optimization with multiple equitable criteria", *RAIRO Oper. Res.*, vol. 33, pp. 275–297, 1999.

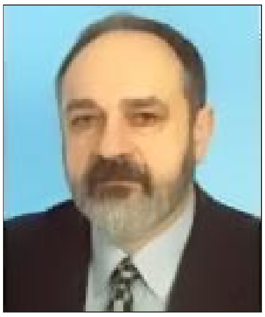
[8] H. Luss, "On equitable resource allocation problems: a lexicographic minimax approach", *Oper. Res.*, vol. 47, pp. 361–378, 1999.

[9] W. Ogryczak, "On the lexicographic minimax approach to location problems", *Eur. J. Oper. Res.*, vol. 100, pp. 566–585, 1997.

[10] W. Ogryczak, "Inequality measures and equitable approaches to location problems", *Eur. J. Oper. Res.*, vol. 122, pp. 374–391, 2000.

[11] J. Rawls, *The Theory of Justice*. Cambridge: Harvard University Press, 1971.

[12] A. Sen, *On Economic Inequality*. Oxford: Clarendon Press, 1973.



Włodzimierz Ogryczak is a professor of Optimization and Decision Support in the Institute of Control and Computation Engineering (ICCE) at the Warsaw University of Technology, Poland. He received both his M.Sc. (1973) and Ph.D. (1983) in mathematics from Warsaw University, and D.Sc. (1997 in computer science from Polish Academy of Sciences. His research interests are focused on theoretical research, computer solutions and interdisciplinary applications in the area of optimization and decision making with the main stress on: multiple criteria optimization and decision support, decision making under risk, location and distribution problems. He has published three books and numerous research articles in international journals. Since 2000 with ICCE, arlier with Warsaw University (Institute of Informatics).

e-mail: W.Ogryczak@ia.pw.edu.pl
Institute of Control & Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Tomasz Śliwiński
e-mail: T.Sliwinski@elka.pw.edu.pl
Institute of Control & Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

A proposition to exploit the partially linear structure of the nonlinear multicommodity flow optimization problem

Paweł M. Białoń

Abstract — Optimization problems arising in telecommunications are often large-scale nonlinear problems. Usually their big size is generated mainly by their linear parts but the existence of small or medium nonlinear parts prevents us from directly tackling them with linear solvers, which are efficient. Instead, the author has proposed a method to decompose big nonlinear problems into nonlinear and linear parts. Its coordination procedure uses two auxiliary solvers: quadratic and pure nonlinear. The procedure falls in the class of projection methods. Special cuts proposed by the author allow to avoid an excessive zigzagging while not enormously increasing the complexity of both the parts. The validity of these cuts can be analyzed within the framework of obtuse cone model. Here the author summarizes the method and analyses its applicability to nonlinear multicommodity flow problems. The structure and particular sizes of this problem make the method useful. The considerations are illustrated by a numerical example with a multicommodity flow problem.

Keywords — multicommodity flow problem, projection methods, large nonlinear problems.

1. Introduction

Nonlinear multicommodity flow optimization problems have become a standard mathematical tool in the areas of networks design and flow control. Unfortunately, such problems are usually large. However, like in many other large nonlinear optimization problems, their large size is formed mainly by linear functions, equations, etc. This big linear part of a large nonlinear problem could be itself tackled with efficient linear programming techniques but one must take into account the existence of the small nonlinear part of the problem. Thus we can only think of solving the problem with the efficiency close to the efficiency with which its linear part alone would be solved.

For this sake the author has proposed in [3] a hierarchical optimization algorithm for large nonlinear optimization problems into a big linear part and a small nonlinear part. The obtained subproblems are: a large quadratic subproblem (with constraints from the linear part of the original problem) and a small nonlinear one (with constraints from the nonlinear part of the original problem). The nonlinear subproblem is computationally easy due to its small size; at least the same applies to the coordination procedure. Thus the efficiency of the whole algorithm depends on the efficiency of the quadratic solver applied to the quadratic subproblem and can reach a very high level due to the observed progress in quadratic programming, polynominal techniques etc.

The original author's proposition was not directed to telecommunication applications; it covered a quite general class of large nonlinear problems with big linear parts. However, four particular structural properties of the problems were needed to make the proposition work properly and efficiently. It turns out that these properties are possessed by nonlinear multicommodity flow (MCF) problems, thus making the proposition especially adequate for these problems. This adequateness is shown in this paper.

The coordination procedure of the author's method is a variant of projection methods for feasibility problems[1] [2, 5, 7] with accelerating cuts. The distinguishing features of the authors proposition are a technique of full cuts cumulation, and specially constructed cuts, so called Z-cuts, that allow to decrease the complication of sets shapes caused by cutting. The initial optimization problem can be reduced to a sequence of feasibility problems with the level control technique [8] and these can be then solved with the author's method.

In Section 2 of this paper the proposed method is first summarized, very briefly and with references to [3]. First, the class of large nonlinear feasibility problems solved by the method is defined. Then the idea of projection methods and accelerating them by cuts are sketched. Then follows the description of the author's proposition, involving: the definition of sets forming the upper-level feasibility problem, the realization of projections with optimization subproblems, the definitions of used cuts and the final algorithm statement.

In Section 3 the MCF problem (of a specific subclass) is defined and the suitability of the proposed method to its solving is indicated. The section ends with a numerical illustration with an artificially created MCF problem, aiming in understanding the proposition. In Section 4 the author gives some conclusions and argues his method can be taken into account as an element in a construction of algorithm solving a large MCF problem.

[1] A *feasibility problem* is a problem of finding a point satisfying a set of constraints.

2. The proposed method

2.1. Large nonlinear problem formulation

The initial optimization task is defined as follows:

$$\begin{array}{ll}
\min\limits_{x\in\mathbb{R}^{n_N}, y\in\mathbb{R}^{n_L}} f(x) & f:\mathbb{R}^{n_N}\to\mathbb{R} \\
\text{s.t. (subject to)} & \\
\tilde{g}(x)\le 0 & \tilde{g}:\mathbb{R}^{n_N}\to\mathbb{R}^{m_N-1} \\
A(x^\top, y^\top)^\top \le b & A \text{ is a matrix of size } m_{LI}\times n \\
B(x^\top, y^\top)^\top = d & B \text{ is a matrix of size } m_{LE}\times n \\
x^{lo}\le x\le x^{up}, y^{lo}\le y\le y^{up}, &
\end{array} \tag{1}$$

where functions f and $\tilde{g}_i$ are continuous, quasiconvex, x^{lo},x^{up},y^{lo},y^{up} are constant vector bounds. The above problem can be reduced [8] to a sequence of feasibility problems $F(Q)$ parametrized with a real number Q. Each problem $F(Q)$ consists in finding $(x^\top, y^\top)^\top$ that satisfies:

$$\begin{array}{c}
g(x)\le 0 \\
A(x^\top, y^\top)^\top \le b \\
B(x^\top, y^\top)^\top = d \\
x^{lo}\le x\le x^{up}, y^{lo}\le y\le y^{up},
\end{array} \tag{2}$$

where function $g:\mathbb{R}^{n_N}\to\mathbb{R}^{m_N}$ was obtained from function $\tilde{g}$ by adding a new coordinate saying how much the goal function value exceeds Q, i.e. $g_i(\cdot)\overset{\text{def}}{=}\tilde{g}_i(\cdot)$, $i=1,\dots m_N-1$, $g_{m_N}(\cdot)\overset{\text{def}}{=}f(\cdot)-Q$.

The feasibility problem has n_N nonlinear variables[2], n_L linear variables, m_N nonlinear inequality constraints, m_{LI} linear inequality constraints, m_{LE} linear equality constraints. Let $m=m_N+m_{LI}+m_{LE}$, $n=n_L+n_N$. The better $m_N\ll m$ and $n_N\ll n$, are fulfilled, the more efficient will be the algorithm.

2.2. The idea of projection methods

Projection methods serve to solving the following *convex feasibility problem*:

Find

$$x\in S\overset{\text{def}}{=}\bigcap_{i=1,\dots m} G_i\,, \tag{3}$$

where $G_i\subset\mathbb{R}^n$ are closed, convex sets. In practice G_i are often defined as sets of points allowed by some constraints. By now we assume that S is nonempty. In the description of the solving process we shall confine ourselves with the case of $m=2$.

For $x\in\mathbb{R}^n$ and a closed convex nonempty $C\subset\mathbb{R}^n$ we shall denote by $P_C x$ the orthogonal projection of x onto C, $P_C x=\arg\min_{y\in C}\|x-y\|^2$. The *projection vector* for such a projection is $P_C x-x$. It can be shown that such a projection is defined uniquely.

The simplest way to search for the solution consists in performing sequential alternate projections onto G_1 and G_2; i.e., given the starting point x^0, we produce a sequence

$$x^1=P_{G_1}x^0,\, x^2=P_{G_2}x^1,\, x^3=P_{G_1}x^2,\ \text{etc.} \tag{4}$$

We assume such projections are easily realizable numerically.

The basic fact in convergence analysis of projection methods is that the projection operator possesses the Fejér contraction property.

Definition 1. *A finite or infinite sequence (x^i) of points in a Hilbert space H has the Fejér contraction property with respect to $C\subset H$ if*

$$\|x^i-c\|^2\ge\|x^{i+1}-c\|^2+\|x^{i+1}-x^i\|^2 \tag{5}$$

for each $c\in C$. Similarly, operator $O:H\to H$ has this property if for each $c\in C$ and $x\in H$ $\|x-c\|^2\ge\|Ox-c\|^2+$ $+\|Ox-x\|^2$.

Fact 1. *Projecting onto a nonempty closed convex set of points in $\mathbb{R}^n$ has Fejér contraction property with respect to this set, and, consequently, to each of its nonempty sets.*

For a proof of the above fact see calculations on page 228 in [11] with $t_{\min}=t_{\max}=1$.

After putting $C=S$ we see that with every projection performed in our algorithm (4) we decrease the squared norm from (any but fixed) point $c\in S$ by at least the square of the appropriate step (projection vector) length. Later it will suffice to assure certain lengths of steps to establish the convergence[3].

Alternatively, the Fejér contraction property of projections in our algorithm means that we approach each solution point with an acute angle.

Zigzagging often slows down projection methods: we may approach the solution with an angle less than but close to $\pi/2$, making the distance from a solution decrease very slowly. This happens in an example in Fig. 1; there, moreover, consecutive projection vectors form angles close to π.

Cuts serve as a standard remedy for zigzagging; a cut is an inequality of the form $\langle\cdot-a,b\rangle\ge\langle b,b\rangle\ge 0$ with fixed $a,b\in\mathbb{R}^n$; *its hyperplane* $H(a,b)$ is given as $\{x\in\mathbb{R}^n:\langle x-a,b\rangle=\langle b,b\rangle\}$, *its halfspace* – as $\{x\in\mathbb{R}^n:\langle x-a,b\rangle\ge\langle b,b\rangle\}$.

Using cuts means replacing (4) with

$$x^1=P_{G_1'^{1}}x^0,\, x^2=P_{G_2'^{2}}x^1,\, x^3=P_{G_1'^{3}}x^2,\ \text{etc.} \tag{6}$$

where sets $G_1'^{k}$ and $G_2'^{k}$ $(k=1,2,3,\dots)$ are G_1 and G_2 narrowed by some cuts, i.e., they were obtained from G_1

[2] A *nonlinear variable* is a problem variable involved in at least one nonlinear function in the model formulation; the remaining variables will be called *linear*.

[3] Which is usually easy and is done with the notion of problem *regularity* [2].

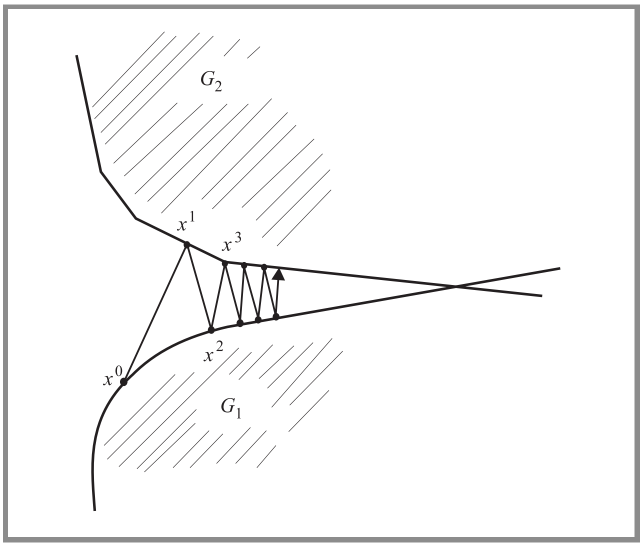


Fig. 1. Zigzagging.

and G_2 by intersecting G_1 and G_2 with halfspaces of some cuts.

A *geometric cut based on (constructed after) the projection* of $x \notin G$ onto close convex G, $G \supset S$ is defined as

$$\langle \cdot - x, P_G x - x \rangle \geq \langle P_G x - x, P_G x - x \rangle.$$

In Fig. 2, unlike in Fig. 1, point x^3 was obtained by projecting x^2 not onto G_2 but onto G_2 narrowed by the geometric cut constructed after projection of x^2 onto G_1. H is a hyperplane of this cut. We see that the step made is longer and we approach the solution with a smaller angle.

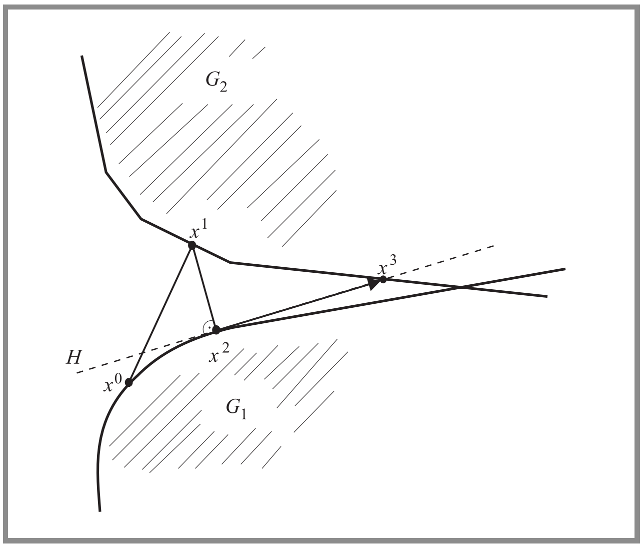


Fig. 2. A geometric cut reduces zigzagging.

A cut is called *valid* or *proper* if it is satisfied for each point in the solution set S. Validity is necessary to assure that projections on narrowed sets (i.e., $G_1'^k$ or $G_2'^k$) still possesses the Fejér contraction property with respect to S; moreover we do not want our method to degenerate by producing empty $G_1'^k$ or $G_2'^k$. Geometric cuts constructed after a projection of an $x \notin G$ onto nonempty, closed, convex $G \supset S$ can be easily shown to be proper.

We may narrow set G_1 or G_2 with only one cut but it may bring a profit in efficiency to narrow them with several cuts simultanously (i.e., to intersect G_1 or G_2 with the intersection of the halfspaces of several cuts). Various techniques for cuts cumulation are given in [4, 5, 9, 10, 13] and a specifically understood cumulation will be also used here.

2.3. The idea of the method

In order to solve our feasibility problem (2) we need to somehow transform it to the form of expression (3).

The following sets N and L will play the role of G_1 and G_2 in (3):

$$N = \{x \in \mathbb{R}^{n_N} : g(x) \leq 0 \wedge x^{lo} \leq x \leq x^{up}\}$$
$$L = \Big\{x \in \mathbb{R}^{n_N} : x^{lo} \leq x \leq x^{up} \wedge \exists_{y \in \mathbb{R}^{n_L}} \Big(y^{lo} \leq y \leq y^{up} \wedge \wedge A(x^\top, y^\top)^\top \leq b \wedge B(x^\top, y^\top)^\top = d\Big)\Big\}.$$

Notice that these are not actually the sets of points allowed by nonlinear and linear constraints but their orthogonal projections on the subspace of nonlinear variables. The projection method will be defined in this subspace.

The projection method of solving the feasibility problem of finding a common point of N and L will form the higher level of decomposition. The lower level will serve to realize the projections.

Finding the projection of point $z \in \mathbb{R}^{n_N}$ onto N may be realized as solving the nonlinear optimization problem

$$\min_{x \in \mathbb{R}^{n_N}} \frac{1}{2}\|x - z\|^2 \quad \text{s.t.} \quad x \in N. \tag{7}$$

Finding the projection of point $z \in \mathbb{R}^{n_N}$ onto L might be realized as solving the quadratic subproblem

$$\begin{gathered}\min_{x \in \mathbb{R}^{n_N}, y \in \mathbb{R}^{n_L}} \frac{1}{2}\|x - z\|^2 \\ \text{s.t.} \\ A(x^\top, y^\top)^\top \leq b \\ B(x^\top, y^\top)^\top = d \\ x^{lo} \leq x \leq x^{up} \\ y^{lo} \leq y \leq y^{up}. \end{gathered} \tag{8}$$

Note that if the solution $(x^{\star\top}, y^{\star\top})^\top$ of the later subproblem satisfies $x^\star \in N$ then it also solves the initial feasibility problem (2). Later one may use either the whole solution $(x^{\star\top}, y^{\star\top})^\top$ of (8) or only vector $x^\star$. The former is appropriate in communication with the user (the printout of final solution) while the later is more convenient in the algorithm description. The following consideration will be in principle done in the subspace $\mathbb{R}^{n_N}$ of nonlinear variables. Placing the higher level of decomposition in this low-sized subspace will certainly increase the efficiency of calculations. This subspace can be really considered low-dimensional when $n_N \ll n$. The other reason why we required $n_N \ll n$ and also $m_N \ll m$ is connected with the nonlinear subproblem (7): both the inequalities make it easy by reducing the number of its variables and of its constraints, respectively.

For the generated sequence of points we shall use the following notation, slightly different from (6) and given in a recursive form:

$$\vec{x}^k = P_{L'^k}\breve{x}^{k-1},\ \breve{x}^k = P_{N'^k}\vec{x}^k \qquad k = 1, 2, \ldots \tag{9}$$

Sets N'^k and L'^k were obtained from L and N by narrowing with some (possibly by no) cuts.

2.4. Cuts

We shall measure zigzagging $Z_{(y^i)}(l)$ (with $k < l$) of a finite sequence $(y^i)_{i=0}^{l}$ of points in a Hilbert space as:

$$Z_{(y^i)}(k) = \frac{\sum_{i=k}^{l-1} \|y^{i+1} - y^i\|}{\|y^l - y^k\|}. \tag{10}$$

If x^i are points generated by some projection method we should try to keep $Z_{\cdot}(i)$ as small as possible. For this sake we shall introduce the *full cumulation* of geometric cuts. Namely, if a cut was constructed after the projection of point x^i onto some set (and x^{i+1} is the result of this projection) then this cut affects all the subsequent projections, which means that these projections are done onto sets narrowed by (maybe not only) this cut. In other words, all the subsequent points x^j must satisfy the cut. One of the alternatives for the full cuts cumulation is using noncumulated cuts: each cut affects only the nearest projections.

For the full cuts cumulation we can nicely assess the sequence zigzagging.

Theorem 1. *Let a sequence $(x^i)_{i=0}^{n}$ (where $n \geq 1$) of points in a Hilbert space satisfies the cumulated geometric cuts condition:*

$$\forall_{s, 1 \leq s \leq n-1}\, (x^s - x^{s-1})^\top (x^n - x^s) \geq 0. \tag{11}$$

Then the following assessment for the sequence zigzagging holds:

$$Z_{(x^i)}(n) \equiv \frac{\sum_{i=0}^{n-1} \|x^{i+1} - x^i\|}{\|x^n - x^0\|} \leq \sqrt{n}. \tag{12}$$

Proof. See the proof of Theorem 1 in [3].

The analysis of the the theorem proof convinces also that usually the zigzagging places below the above limit; inequality (12) is fulfilled as equality only for very particular configurations of points x^i.

We shall describe the firts two types of cuts present in the method. In kth iteration the following cuts are constructed:

1. $\langle \cdot - \vec{x}^k, \breve{x}^k - \vec{x}^k \rangle \geq \langle \breve{x}^k - \vec{x}^k, \breve{x}^k - \vec{x}^k \rangle$ – type A cuts. They are later used, once or many times[4], to narrow set L.
2. $\langle \cdot - \breve{x}^k, \vec{x}^k - \breve{x}^{k-1} \rangle \geq \langle \vec{x}^{k-1} - \breve{x}^{k-1}, \vec{x}^k - \breve{x}^{k-1} \rangle$ – type B cuts. They are later used, once or many times, to narrow set N.

These are simply geometric cuts, but we distinguish the cuts made after projections onto N'^k (type A) and after projections onto L'^k (type B).

It is possible to apply Theorem 1 to our algorithm. If we take sequence $\breve{x}^0$, $\vec{x}^1$, $\breve{x}^1$, $\vec{x}^1$, ... as sequence (x^i) in this theorem, cumulating both A-type and B-type cuts will assure the satisfaction of (11) for $n \geq 1$, thus the sequence will not zigzag too strongly.

However, the cuts of both the types have their numerical drawbacks that increase in case of cumulation. A-cuts influence the definition of (subsequent) sets L'^i and thus complicate the quadratic optimization subproblem. The complication may consist in introducing nonzero elements in the sparse constraint matrix of this problem (approximately n_N ones per cut). Also, we cannot be certain that the subsequent cuts will not decrease the problem conditioning, e.g. by aligning almost in parallel. The main disadvantage of cumulating B-cuts origins from the small size of the nonlinear optimization subproblems: the relative complication introduced in these problems by many cuts may be large.

Fortunately, it has turn out possible to resign cumulating cuts of one of the above types in the algorithm, while preserving the applicability of Theorem 1. With the trick described later, the user may resign generating (not only cumulating!) cuts of one of the types. The choice of the type should depend on particular problem properties. Due to the symmetry of the question, from now we shall only consider the case of giving up generating the B-type cuts.

The trick consists in generating in kth iteration cuts of the third type:

3. $\langle \cdot - \vec{x}^k, \vec{x}^k - \vec{x}^{k-1} \rangle \geq \langle \vec{x}^k - \vec{x}^{k-1}, \vec{x}^k - \vec{x}^{k-1} \rangle$ – type Z cuts. They are later used to narrow set L.

When we take sequence $(\vec{x}^i)$ as sequence (x^i) in the assumption of Theorem 1 and decide to cumulate Z-cuts, the theorem will limit the zigzagging of this sequence. However, we must prove that each Z-cut is proper. Fortunately, we can show the propriety of the Z-cut constructed in kth iteration, on condition that the A-cut constructed in $(k-1)$th iteration was taken into account in definition of L'^{k-1}. This Z-cut is shown to be proper as implied by two proper cuts: the mentioned A-cut constructed in iteratin $k-1$ and the B-cut that we might have (but have not) constructed in iteration k (see Fig. 3).

[4] Depending on our decision about cumulating the cuts.

Showing this implication exceeds the scope of this paper (see Theorem 2 in [3]). In the proof the conical cuts surrogating method [4, 5, 9] were used. Surrogating techniques enable showing the propriety of a certain constructed cuts (so called surogate cut) from the propriety of several other cuts.

The applied trick is also similar to the modification in Section 5 of [4].

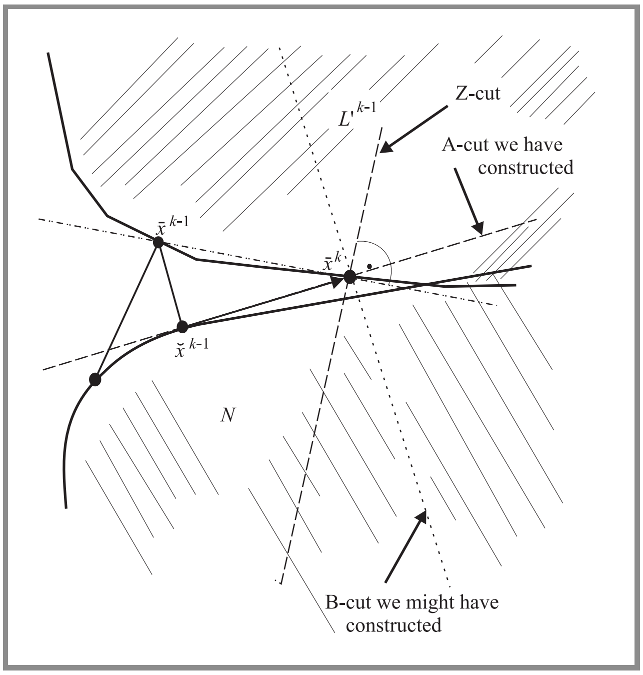


Fig. 3. Construction of Z-cut in kth iteration.

2.5. The algorithm

The algorithm will be given in its basic variant, in which B-cuts are absent, A-cuts are not cumulated, Z-cuts are cumulated.

Algorithm 1. **Parameters:** tolerance $t^N \geq 0$, starting point $\breve{x}^0$.
We initialize the iteration counter k with 1.

1. Compute $\vec{x}^k = P_{L'^k}\breve{x}^{k-1}$ with L'^k being L narrowed by some cuts constructed in earlier iterations:

$$L'^k = \{y \in L : \langle y - \vec{x}^{k-1}, \breve{x}^{k-1} - \vec{x}^{k-1}\rangle \geq \\ \langle \breve{x}^{k-1} - \vec{x}^{k-1}, \breve{x}^{k-1} - \vec{x}^{k-1}\rangle \wedge \\ \wedge (\forall_{j \in K^k} \langle y - \bar{x}^{j-2}, \bar{x}^{j-1} - \bar{x}^{j-2}\rangle \geq \\ \langle \bar{x}^{j-1} - \bar{x}^{j-2}, \bar{x}^{j-1} - \bar{x}^{j-2}\rangle\},$$

   where K^k equals to $\{3, \ldots k\}$, by solving the quadratic subproblem (8) with L replaced by L'^k and with the substitution $z \leftarrow \breve{x}^{k-1}$. If $L'^k = \emptyset$ then STOP – report infeasibility.

2. Compute $\breve{x}^k = P_N \vec{x}^k$ by solving the nonlinear subproblem (7) with the substitution $z \leftarrow \vec{x}^k$. If $N = \emptyset$ then STOP – report infeasibility. If $\|\breve{x}^k - \vec{x}^k\| \leq t^N$ then STOP – return the last solution of the quadratic subproblem. Otherwise set $k := k + 1$ and go to step 1.

A detailed convergence analysis of (a slightly more general) method is given in [3, Section 7]. It bases on Fejér contraction property and the regularity analysis of the problem and zigzagging; since the used cuts are proper, the Fejér contraction mechanism is not disturbed. The analysis conceives also the case of infeasibility: $L \cap N = \emptyset$. Based on the guaranteed sequence zigzagging, the moment of detection of infeasibility is assessed.

3. The applicability of the method to the multicommodity flow problem

3.1. The multicommodity flow problem

We shall formulate a variant of a problem of the well known class of multicommodity flow problems [15]. Let us represent a telecommunication network as a directed graph. Let the graph nodes be represented by integer numbers from the set $I \overset{\text{def}}{=} \{1, \ldots, N\}$, the directed arcs – as members of a set $E \subset I \times I$ (arc (i, j) will correspond to the unidirectional link from node i to node j).
Various commodities (various kinds of information) are to be transported through our network; let us number them with $1, \ldots K$. The demand on kth commodity in ith node is given by the real parameter $r^{i,k}$, while its negative value denotes that the node actually emits the commodity (in the amount of $|r^{i,k}|$). Define decision variables in our problem:

- $\phi_u^k \in \mathbb{R}$, $u \in E, k = 1 \ldots K$ – the kth commodity flow in arc u.
- $\psi_u \in \mathbb{R}$, $u \in E$ – total flow of all commodities in arc u.

The flow ψ_u in arc u costs $\Phi(\psi_u)$, where $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is an increasing function. The cost can have various real-world interpretations. For example, it can represent the cost of reconstruction of link u to the capacity of ψ_u or it can be a certain measure of slowness of the link.
The multicommodity flow optimization problem consists in minimizing the total cost of network flow and is formulated as follows:

$$\min_{\phi_u^k} \sum_{u \in E} \Phi(\psi_u) \tag{13}$$

$$\text{s.t.}$$

$$\psi_u = \sum_{k=1}^{K} \phi_u^k \quad \text{for } u \in E \tag{14}$$

$$\sum_{(i,j) \in E} \phi_{(j,i)}^k - \sum_{(i,j) \in E} \phi_{(i,j)}^k = r^{i,k} \text{ for } i \in I, k = 1, \ldots K \tag{15}$$

$$\phi_u^k \geq 0 \quad \text{for } u \in E, k = 1, \ldots K. \tag{16}$$

Equation (13) defines the total cost of network flow, (14) defines the total flow in each arc u, (15) expresses the Kirchoff law for each node i and, finally, (16) reflects the unidirectional character of arcs.

The resulting feasibility problem $F(Q)$ takes the form

$$\begin{aligned} &\text{Find } \phi_u^k \in \mathbb{R}, \psi \in \mathbb{R} \\ &\text{satisfying} \\ &\sum_{u \in E} \Phi(\psi_u) \le Q \text{ and} \\ &(14) - (16). \end{aligned} \tag{17}$$

3.2. The applicability of the proposed method

Let us summarize the important properties of the feasibility problem (2) important for the proper and efficient work of our algorithm:

1. The nonlinear equality constraints are absent.
2. The nonlinear inequality constraint functions must be quasiconvex.
3. $n_N \ll n$, where n_N is the number of nonlinear variables, n is the number of all variables, is appreciated.
4. $m_N \ll m$, where m_N is the number of nonlinear constraints, m is the number of all constraints, is appreciated.

The key observation in this paper is that these properties are possessed by problem (17):

Ad 1. Obviously.

Ad 2. The increasing character of Φ implies its quasiconvexity and thus the quasiconvexity of $\sum_{u \in E} \Phi(\psi_u)$ treated as a vector function of ψ_us. It should be stressed that continuous, increasing but concave Φ, typical in practice due to the economy-of-scale phenomenon, are acceptable[5].

Ad 3. Note that $n_N = |E|$, $n = K \cdot |E|$.

Ad 4. Note that $m_n = 1$, $m = K \cdot N$.

Observe also that the last two properties are the better fulfilled the greater is the number K of commodities.

3.3. Numerical illustration

The method was applied to an artificial multicommodity flow problem of class (13)–(16). The aim of experiments was to show the relations between particular sizes of the problem (and subproblems) we can deal with, to simply validate the method by analyzing its results and to investigate how much iterations do the coordination procedure of our method as well as the level control loop (the costs of optimization of subproblems were not investigated, since they depend on many technical details: used solvers, restarting techniques, etc.).

The bidirectional ring network, shown in Fig. 4 was used in computations.

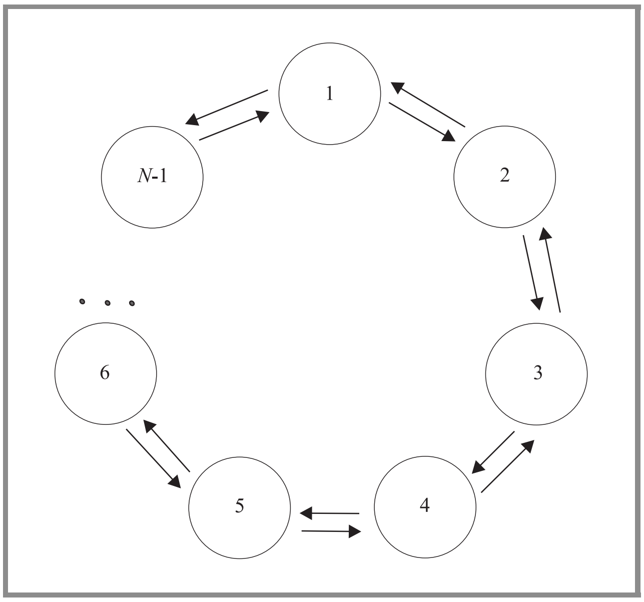


Fig. 4. The example network. Circles represent nodes, arrows represent links, numbers in circles – numbers of nodes.

$K = N$ commodities were distinguished. Each kth commodity had a single source node, namely node k, and a single collector node, namely node $((k+1) \mod N) + 1$, so each commodity flew clockwise between two consecutive nodes. The flow of kth commodity amounted to the value of $1.5 \cdot k/N$. Precisely, there was:

$$\begin{aligned} &r^{k,k} = 1.5 \cdot k/N: && \text{for } k = 1, \dots N \\ &r^{k,((k+1) \mod N)+1} = -r^{k,k} && \text{for } k = 1, \dots N \\ &r^{i,j} = 0 && \text{for remaining pairs } (i, j). \end{aligned}$$

The number of nodes N was the parameter of the problem, and the problem structure implied the remaining sizes: $|E| = 2$ and, as said above, $K = N$.

Table 1 shows the particular sizes of the problem, seen as an instance of optimization problem (1). The same sizes are adequate also for the resulting problem (2). The number m_N of nonlinear constraints equals 1. The sizes of optimization subproblems from the decomposition scheme can be also reconstructed from this table: the nonlinear subproblem has n_N variables and $m_N = 1$ constraints and the quadratic subproblem has n variables and m constraints.

The cost function Φ was defined as $\Phi(\psi) = (1 + \psi^2)^{0.4} - 1$. For small arguments this function behaves like a convex function, whereas for large arguments – like a concave one. Such a choice was aimed to show the broadness of the class of functions Φ acceptable by our method; also it introduces the speciffic phenomena in the optimized network flow (see later). It can have the following real-world interpretation: the cost of reconstruction of a link should be in principle

[5] It remains to explain why our method required property 2. It was simply necessary to make the level sets of the constraint functions convex and thus the projection methodology applicable.

given by a concave function, to acomodate the economy-of-scale phenomenon. However, for small flows, a serious reconstruction of a link might be not necessary, it perhaps suffices to make small improvements. Thus for small arguments Φ should be rather flat.

Table 1
The problems and their sizes

Problem	A	B	C	D
N	3	10	20	30
Number of variables (n_N)	25	220	840	1860
Number of constraints (m)	15	120	440	960
Number of nonlinear variables (n_N)	6	20	40	60

The reduction of optimization problem (1) to a sequence of feasibility problems (2) was done with the level control scheme [8]. This method can be viewed as a method for finding the optimal value of the problem (1). It is based on bisection of a certain interval. The initial left end L and right end U of the interval are given by the user: L and U are lower and upper bounds for the optimal value. The value Q for current feasibility problem Q is chosen somewhere in the midle on the current interval. The interval is narrowed with the following techniques:

- Values of f in feasible points generated during the algorithm course are used to update the lower bound (left end of the interval).
- Infeasibility of the feasibility problem $F(Q)$ allows to update the current upper bound (right end of the interval) to the value of Q. The infeasibility is detected by encountering that the sum of squares of made steps exceeds the square of R, the user-given diameter of a ball containing all the points generated by the algorithm[6].

The applied method varied from the original method of level control in the following aspects:

- The cuts were present when solving the feasibility problems.
- Infeasibility of a feasibility subproblem was detected much quicker by encountering the emptyness of N or L'^k (which, in turn, was detected as an infeasibility of one of the optimization subproblems), similarily as in [6].

The simple structure of the problem allows to quess its optimal value (the minimal cost). Each commodity can be reasonably sent between its source and its collector (the consecutive nodes) only in two ways: clockwise (through a single link) or clock-counterwise (through a path of length $N-1$). Since the later way engages much more links, it probably generates a bigger cost. Thus we can transport all the commodity clockwise. For such a solution the total flow cost will be certainly equal to $\sum_{k=1}^{N}\Phi(r^{k,k})$, and it will be refered later as the *heuristic optimal value of the problem or heuristic* $f^\star$, since we obtained it with a heuristic reasoning.

The method was implemented in the C++ language. The LP-DIT library [14] implementing sparse matrices and realizing linear problems storage was used.

The solver from IAC-DIDASN++ system (see e.g. [12]) and HOPD [1] were used as auxiliary solvers: respectively nonlinear and quadratic.

The parameters of the level control scheme were set as follows: $L=0$, $U=5N$, $R=50N$, $\theta=0.75$ (θ is a bisection parameter – see op.cit.). Tolerance t^N was set to $1e-4$. The results of experiments are given in Table 2, where $f^\star$ denotes the optimal value of a problem, "nfp" – the number of feasibility problems generated by the level control scheme and "total iterations" – the total number of iterations the method did in solving all feasibility problems.

Table 2
Results of experiments

Problem	A	B	C	D
Heuristic $f^\star$	1.52605	5.76155	11.81417	17.86702
Computed $f^\star$	1.37162	5.73391	11.8015	17.8588
nfp/total iterations	9/34	12/31	11/31	12/26

Experiments are commented in Section 4, here we shall only show why the comuted optimal values seem reasonable. They are slightly lower than their "heuristic" variants. This explains in the following way. While the flow in clock-counterwise is very small, the derivative of Φ at the point corresponding to such a flow is also very small. Thus it pays to send a small fraction of each demand a clock-counterwise-way, which is cheaper and which we did not take into account in our heuristic reasoning. The analysis of the values of decision variables obtained by the method supports this hypothesis. However, with the grow of N, the length of the clock-counterwise path becomes larger, the costs of sending flows clock-counterwise – larger, and the described phenomenon vanishes. This we see in Table 2: the gap between the heuristic and computed optimal values clearly vanishes with the growth of N.

4. Conclusions

The goal of the author was to show that multicommodity flow problems are a very interesting case of large nonlinear optimization problems that seem especially created for his method, mainly due to their particular sizes and absence of nonlinear equality constraints.

[6] Which, roughly speaking, contradicts to the behavior implied by the Fejér contraction property.

The simple example numerical presented above was only an illustration. It does not conceive the complexity of practical models, with their additional structural elements, some hierarchical structure necessary to account for extremely huge sizes, etc. However, the basic information for a constructor of a software solving MCF refers to the behavior of the projection method itself and is following: the method did not do many iterations, and their number did not grow with the increase of the problem size. Moreover, the size of the subspace in which the projection method operates, as well as the sizes of nonlinear subproblems were really small. Thus embedding the method in such a software thus seems worth considering. However, many technical details ought to be dealt with. The first one will perhaps concern warm restarts. Optimization subproblems, especially quadratic ones, are very similar each to other and should not be solved independently, but in each subproblem some information (e.g., some matrix factorization) from an earlier instance of the subproblem, should be preserved in order not to repeat similar computations. However, not all the quadratic solvers allow for warm restarts (e.g., the version of the quadratic solver used by the author).

Using projection methods in the presented way gives also some light to the question of how big the complication introduced to big linear MCF problems by the addition of a small nonlinear part is. This complication expressed here mainly with a number several tens of iterations in the higher level of our decomposition, and with the necessity of taking into account quadratic goal functions. Both these aspects of complication may turn out to be modest by the current and future growth of computers power and progress in quadratic programming techniques; moreover, techniques like warm restarting can decrease their meaning.

Finally, we state that, despite the precise convergence analysis given in [3], the simple example showed the validity of the method in the sense of finding a proper solution.

References

[1] A. Altman and J. Gondzio, "Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization", *Optim. Meth. Softw.*, vol. 11–12, pp. 275–302, 2000.

[2] H. Bauschke and J. Borwein, "On projection algorithms for solving convex feasibility problems", *SIAM Rev.*, vol. 38, no. 3, pp. 367–426, 1996.

[3] P. Białoń, "Large-scale nonlinear projection algorithm using projection methods", *Discus. Math. Differ. Inclus., Contr. Optim.*, vol. 20, no. 2, pp. 171–194, 1999.

[4] A. Cegielski, "A method of projection onto an acute cone with level control in convex minimization", *Math. Progr.*, vol. 85, pp. 469–490, 1999.

[5] A. Cegielski, *Relaxation Methods in Convex Optimization Problems, Higher College of Engineering*. Series Monographies, no. 67. Zielona Góra, Poland (in Polish).

[6] R. Dylewski, "Numerical behavior of the method of projection onto an acute cone with level control in convex optimization", *Discus. Math. Differ. Inclus., Contr. Optim.*, vol. 20, 1999.

[7] S. Flåm and J. Zowe, "Relaxed outer projections, weighted averages and convex feasibility", *BIT*, vol. 30, pp. 289–300, 1999.

[8] S. Kim, A. Hyunsil, and C. Seong-Cheol, "Variable target value subgradient method", *Math. Progr.*, vol. 49, pp. 356–369, 1991.

[9] K. Kiwiel, "Monotone Gram matrices and deepest surrogate inequalities in accelerated relaxation methods for convex feasibility problems", *Linear Algeb. Appl.*, vol. 215, pp. 27–33, 1997.

[10] K. Kiwiel, "The efficiency of subgradient projection methods for convex optimization", Part I: "General level methods", *SIAM Contr. Optim.*, vol. 34, no. 2, pp. 660–676, 1996.

[11] K. Kiwiel, "Block-iterative surrogate projection methods for convex feasibility problems", *Linear Algeb. Appl.*, vol. 15, pp, 225–259, 1995.

[12] T. Kręglewski, J. Granat, and A. Wierzbicki, *IAC-DIDAS-N – A Dynamic Interactive Decision Analysis and Support System for Multicriteria Analysis of Nonlinear Models, v. 4.0.* Laxenburg (Austria): International Institute for Applied Systems Analysis, June 1991, collabor. paper, CP-91-101.

[13] C. Lemaréchal, A. S. Nemirovskii, and Yu. Nesterov, "New variants of bundle methods", *Math. Progr.*, vol. 69, pp. 111–147, 1995.

[14] M. Makowski, *LP-DIT Data Interchange Tool for Linear Programming Problems (Version 1.20).* Laxenburg (Austria): International Institute for Applied Systems Analysis, 1994, work. paper, WP-94-36.

[15] M. Minoux, "Network synthesis and optimum network design problems: moodels, solution methods and applications", *Network*, vol. 19, pp. 313–360, 1989.

[16] M. Shchepakin, "On a modification of a class of algorithms for mathematical programming", *Zh. Vychisl. Mat. i Mat. Fiz.*, vol. 19, pp. 1387–1395, 1979 (in Russian).

Paweł M. Białoń was born in Warsaw in 1971. He received his M.Sc. in computer science from Warsaw University of Technology in 1995. Currently he is employed by National Institute of Telecommunications in Warsaw. His research focuses on nonlinear optimization methods and on decision support and is directed towards the Ph.D. degree. He has participated in several projects applying decision support methods in the telecommunications, agricultural and environmental areas.

e-mail: P.Bialon@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paper

Decision support tool for web cache management

Jarosław Pietrzykowski

Abstract — Web caching is the subject of intense research and development since it seems to be very promising area. Web caching means storing copies of frequently used objects (documents) geographically close to users requesting them to reduce network load, servers load and user response times. Cache can be situated in different locations between user and servers with original content. It seems that the most significant improvement can be achieved by using the proxy server – a dedicated web server. Various parameters affect web cache performance: cache size, limitations on document sizes or documents removal policy. Several metrics are used to evaluate this performance: hit rate for example is the ratio of documents obtained by using the cache mechanism comparing to the total number of documents requested. Choosing adequate parameters for interesting traffic patterns to improve cache performance is not a trivial cache management problem. Prototype tool based on multicriteria model analysis and supporting such web cache management is presented. Simple case was examined where small number of sets of cache parameters (variants), miniature traffic representation and only two performance measures are considered.

Keywords — web cache management, multicriteria analysis.

1. Introduction

In the era of immense grow of Internet advanced solutions for network overload reduction have special importance. One of such solutions is web cache acting as a proxy between users and web sites. This paper concentrates on some issues connected with web proxy cache management. Problem considered here concerns choosing configuration parameters for web cache for a given requests stream representation in order to improve cache performance.

Suggestions for using cache parameters are general and does not necessarily apply to individual web cache placed in specific network environment. Such environment consists of population of computer users and computer infrastructure providing connectiviness between users and access to external network, like Internet. Users activity induces specific traffic for their network. This traffic commonly does not bear constant characteristic. Some patterns emerge repeatadly and some new can be observed. Experienced web cache managers are able to tune parameters mechanism accordinlgy to changes in traffic directed to the cache. Nevertheless assesment of the influence of cache parameters on its perfomance becomes harder with each new version of the software because number and complexity of parameters increases. Neccessity of choosing from only dozen of configuration alternatives can be very confusing and human intuition can fail.

This paper presents concept and its realization of decision support tool that could help web cache managers with choosing adequate configuration for better cache performance. Such tool should conform to several requirements. It should allow for examination many configuration alternatives for interesting patterns of requests stream flowing into the cache. During evaluation of cache performance for these alternatives several criteria should be regarded. Important feature of software supporting decision making is to enable clear and easily understandable expression of decision maker preferences concerning used criteria. It is crucial for the correct measurement to separate examined cache from unstability of external network and ensuring that experiment is repeatable. Analysis of cache configuration alternatives should not disturb regular work of cache software. Such analysis should also allow for fast evaluating many alternatives. High level of automation and graphical, user-friendly interface are desirable.

In short perspective development of such tool complying with presented requirements was the main target of the work presented here. This includes inventing detailed concept of the tool, creating its components and testing them separately and assembled together and also examining the whole on some real example. This stage of work presented in this paper was intended also for gaining better acquaitance with web cache subject, capabilities of this category of computer software and for identification of specific problems. Some important results of this work are software component generating mathematical model for considered decision problem and two-phase design of performance measurement operation so that appropriate requirements are satisfied.

In the future more advanced system for automatic web cache configuration is planned. Such system will recognize traffic patterns in requests stream flowing into the cache and will apply corresponding set of parameters in order to obtain best performance. The role of the tool presented in next part of this paper will be to match configurations with new patterns for future use by this automated system.

Following chapter contains simple description of web cache concept and associated issues. Then details of the tool supporting tuning of web cache parameters and outline of its advancement to automated system are presented. Next, application of the tool is illustrated with some simple case. Finally short summary follows.

2. Web cache concept

2.1. Internet traffic explosion

Tremendous increase in Internet traffic has been observed recently, especially with regard to World Wide Web area [1]. This results in servers overload, congested networks and delays in response times of document requests. This trend is augmented by fast development of new Internet applications such as electronic commerce, business-to-business exchange and multimedia communication. There is a huge demand for data, information and knowledge concerning almost all human activities.

Apparently, this demand cannot be satisfied by advances in Internet infrastructure development like high capacity backbone networks, cable modems or radio access. Other approaches are neccessary that apply more "soft" techniques like server load balancing, better traffic management or web caching.

Web caching basically means that copies of popular WWW documents (such as html pages, graphics or audio files – altogether called objects) are kept geographically close to users requesting them. When document request comes to web cache its mechanism tries to handle it by retrieving desired document from its own storage space. In the case of success document is delivered immediately and "hit" is recorded. Otherwise requested object has to be downloaded from external network, usually from the server where original document resides. In such situation web cache "miss" is recorded. Cache parameters and document's features determine if this document is stored in web cache for future use.

The idea of caching is not new and has been widely used in computer science area. Applications of this mechanism can be found both at hardware and software level. Usually it causes significant increase in performance. Recent incredible advance in microprocessor design partially results from sophisticated use of this concept.

2.2. Benefits of using web cache

Advantages of using cache depend on its location within network structure. Basically there are three possible places on the way from user to server with desired content where cache is situated:

1) user's machine – **client cache**;

2) original server – **server cache**;

3) in the middle of network – **proxy cache**.

These three types of web cache are shown in Fig. 1.

In the first case user of client machine benefits most from applying cache. But still if many users need access to same document corresponding number of objects must be downloaded from the original server.

When web cache is situated on the side of server with original content this does not decrease network overload significantly. Nevertheless it reduces load on the server and shortens response time to the document request.

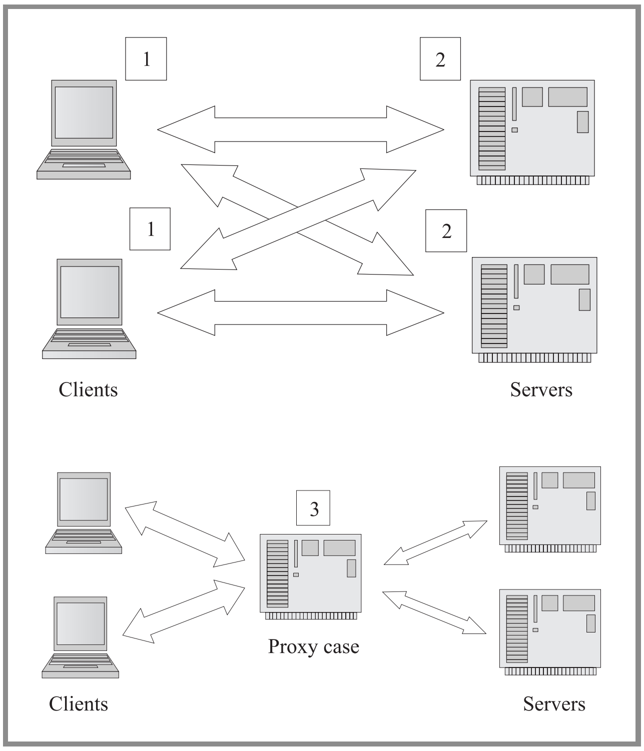


Fig. 1. Proxy cache placement in the client-server network.

Proxy web cache is the most fruitful solution. Response time is reduced considerably because it is kept closer to the requester. Additionally, if document is popular among other users it results in substantial decrease in network congestion since number of requests is much smaller. This approach also causes reduction of original server load. This is especially important when Internet access is expensive. Proxy caches can make up hierarchies, where bigger ones handle smaller ones' requests or they can be organized as peer nodes network. There are some large national proxy cache networks which are used intensely by scientific society. SURFnet in The Netherlands or UNINETT in Norway can suit as an example (these two are connected to each other). Savings resulting from applying this solution are reported to be 30% to 50% of the traffic, depending on the traffic characteristic[1]. At all levels of such hierarchy advantages prevail drawbacks and significant decrease in response time is noticed. Although benefits from using web cache are the most apparent in large organizations it is recognized that such mechanism is useful also in small scale. Web proxy cache software finds also other applications within organization, such as:

- incoming traffic filtering;
- accelerating access to overloaded servers;

[1] http://www.desire.org/html/services/caching/

- protection against computer assaults from external network.

Web cache proxy mechanism is one of the most important solutions providing improvement in network utilization efficiency. It has been area of intense research for new algorithms, traffic models and other concepts leading to extending set of parameters designed for better performance. In the following parts of this paper term "web cache" means its proxy type since it is more concise form and can be used without confusion.

2.3. Performance measures

Web cache performance can be measured in several ways. The most popular metrics are:

- **document hit ratio** – defined as ratio between number of documents delivered with the use of web cache mechanism and total number of requests;
- **byte hit ratio** or **weighted hit ratio** – amount of bytes retrived from network using cache compared to total amount of bytes requested;
- **average user response time**;
- **bandwidth utilization**.

There is a trade-off between first two metrics. Often web cache load analyses show that most of requests pertains to small documents. Therefore, in order to achieve high document hit ratio large number of smaller documents should be kept in cache. On the other hand these analyses demonstrate that transfers of huge documents increase network traffic significantly. So as to obtain high byte hit ratio several large documents should be stored in cache at the expense of smaller ones. Importance of these metrics depends on situation:

- when response time reduction is crucial – document hit ratio should affect cache configuration decisions mostly;
- when saving bandwidth is critical – cache configuration should result in high byte hit ratio values.

There are factors influencing web cache performance that are beyond managing person's control. Among these factors are some documents features determined by their authors that prevent them from caching or that set time limit for keeping such documents in cache with regard to their up-to-dateness. Usually authors or owners of web documents want to avoid caching because of possible decrease in their profits. Profitability of many web sites depends on count of accesses to presented web pages (documents). When pages are kept in proxy cache one cannot be sure if their download counts are properly reported to original server. Security concerns are also one of reasons against caching. Documents stored on unknown machine somewhere in Internet cannot be controlled sufficiently by their owners and danger of content abuse is real.

2.4. Configuration parameters

Modern web cache software is complex. Squid - the most commonly used application has more than 150 configuration parameters. These parameters cover many aspects of web cache functionality including network setup, timeouts, access rights and other. Many parameters affect web cache performance significantly and allow controlling the way this mechanism works. Among them are:

- amount of computer memory dedicated for documents storage;
- disk space amount dedicated for caching;
- document replacement policy – determines how many and which documents are swept out from storage space when there is no room for freshly retrieved ones;
- maximal size of kept documents;
- minimal size of kept documents – both parameters allow traffic filtering;
- swap upper threshold (or **cache swap high**) – this parameter is algorithm-specific and means level when documents begin to be swept out more aggressively;
- swap lower threshold (or **cache swap low**) – when this level is approached process of removing documents from cache starts;
- method of stored content refreshment:
  - **passive caching**: documents are stored in cache only if it is requested by client machine;
  - **active caching**: freshness of documents is checked by cache itself which is useful option for popular but quickly outdateing documents;
- rules for document refreshing – specify when document is considered fresh.

Very important parameter is the cache removal policy. Software used in the research presented here implemented three types of algorithms for this task:

- LRU (last recently used) – documents not requested for the longest period are removed;
- GDSF (greedy dual size frequency) – removal is performed on the basis of sizes of documents and cost of their retrieval from the external network;
- LFUDA (least frequently used with dynamic ageing) – documents that have been requested less frequently and of certain age are removed.

Web cache performance is an area of many studies. This results in still extending set of cache parameters. For example the last two algorithms stated above were developed in Hewlett-Packard laboratories and were shortly after implemented in another version of web cache software.

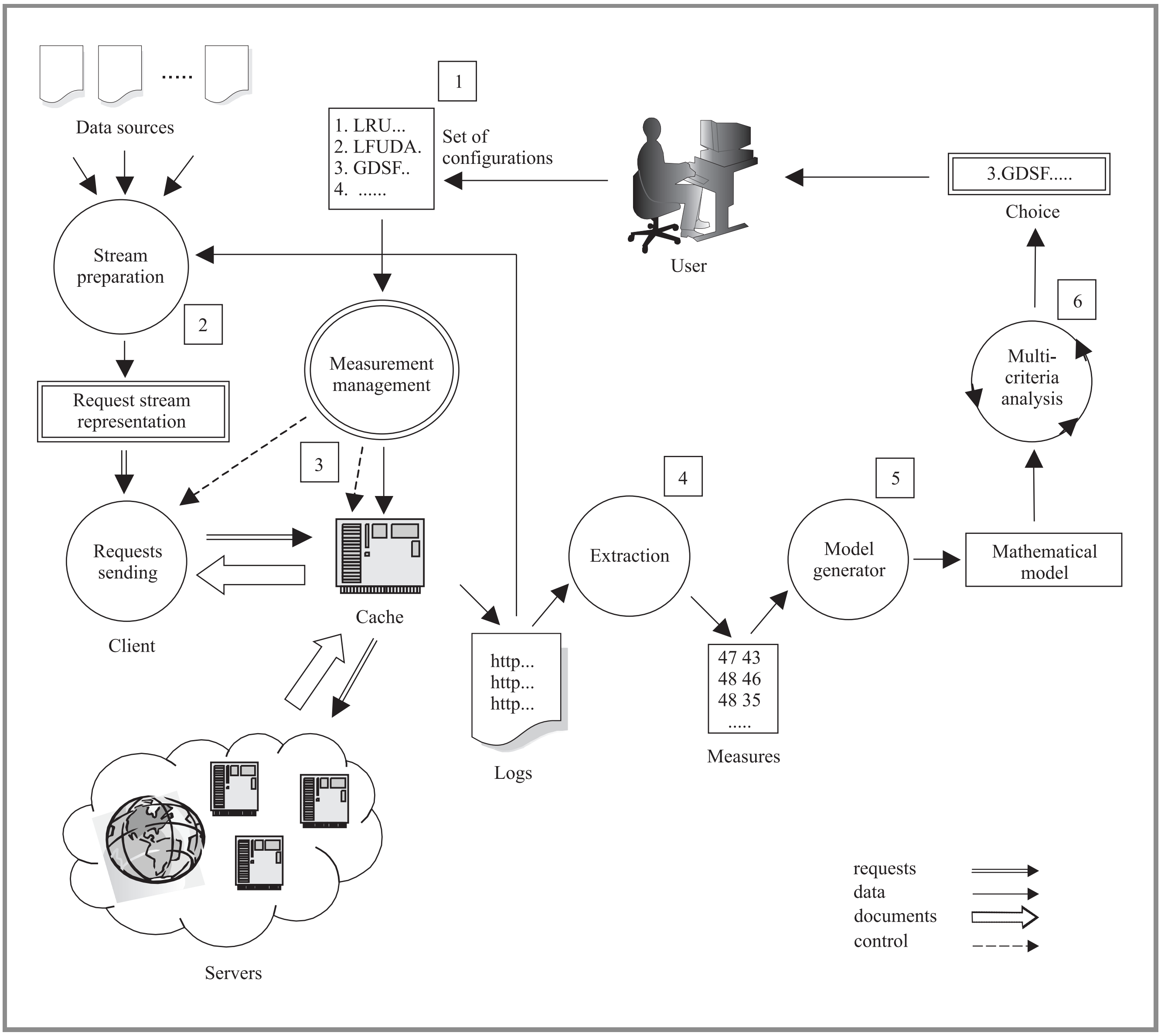


Fig. 2. Process of analysis of decision alternatives for proxy cache configuration.

3. Decision support tool concept

In order to meet requirements presented earlier the tool supporting web cache management should be able to perform several tasks. These tasks should be organized in a process that leads to reasonable choice about better configuration for examined cache. Figure 2 presents such process. In this process, for a given representation of stream of documents requests a number of possible sets of proxy cache parameters are examined. Each set of parameters (configuration) can be seen as a decision alternative. Collected performance results are used to evaluate these alternatives. Multicriteria analysis methodology based on mathematical modeling is used for the purpose of analyzing decision alternatives. The effect of such analysis is the best – according to preferences of person performing analysis – configuration, that can be matched with considered requests stream representation.

The process illustrated in Fig. 2 consists of following stages:

1) preparation of configuration alternatives;

2) preparation of requests stream representation;

3) performance measurement;

4) extraction of output variables;

5) mathematical model generation;

6) multicriteria problem analysis.

Stages 2 and 3 are coupled because of requirements presented earlier that performance measurement should meet.

The details of this process are presented in the following parts of this chapter.

At present the elements of the process constitute rather some analytical environment than fully integrated tool. Nevertheless valuable analyses of web cache configuration alternatives for interesting requests patterns are possible. Author hopes it will be helpful for web cache managers to improve web cache performance.

The process being described here has been implemented as several connected components responsible for performing individual stages mentioned above. Some of the components are UNIX shell scripts, some are realized as programs written in C++ language and some of them are ready to use pieces of software. The latter enclose ISAAP (interactive specification and analysis of aspiration-based preferences) modular tool and MOMIP (modular optimizer for mixed integer programming) solver, which are used for multicriteria model analysis, and WGET program used for sending prepared requests to web cache.

Although some of the stages – like performance measurement and extraction of output variables – were automated, there are still some tasks (namely: preparation of configuration alternatives and preparation of requests stream representation) that has to be done manually or with use of other tools. These parts need more automation so that they could be seamlessly integrated with other elements into one tool. In the future such tool should coordinate performed tasks and provide user-friendly graphical interface. X Windows or Java environments seem to be quite suitable for this purpose.

The software used in this research is distributed either freely as an open-source code (web cache) or with the GNU license (request sending client) or free of charge for non-commercial and educational purposes (multicriteria analysis tool and solver).

All experiments were planned for and conducted with appliance of Squid[2] software (version 2.3.STABLE4) running on Sun Solaris platform. Although there are WWW servers with built-in cache functionality (i.e. Apache) Squid is dedicated for caching. Thus its code is less complicated and more reliable. Squid is also most widely used software in its category and has been awarded several times.

3.1. Preparation of configuration alternatives

This task is performed manually. During this stage a file describing the plan for the experiment has to be prepared in a special format (Fig. 3). In this file there are two sections:

- # para – this section contains in each row names of the cache parameters and number of their values that are to be used in the experiment;
- # data – includes combinations of values of examined parameters.

```
#para
replacement_policy 3
cache_swap_low 3
cache_swap_high 2
#data
LRU 80 95
LFUDA 80 95
GDSF 80 95
LRU 85 95
LFUDA 85 95
GDSF 85 95
LRU 90 95
LFUDA 90 95
GDSF 90 95
LRU 80 100
LFUDA 80 100
GDSF 80 100
LRU 85 100
LFUDA 85 100
GDSF 85 100
LRU 90 100
LFUDA 90 100
GDSF 90 100
```

Fig. 3. Example of configuration alternatives specification.

Preparing only several alteratives of cache configurations can be done manually but there can be demand for dozens to be examined. This can happen quite often since when dealing with only few parameters hundreds combinations of their values make up possible alternatives. It is also possible that some of this combinations are not interesting for cache manager and should be excluded. This leads to the conclusion that this stage must be automated in order to make the presented web cache management support tool truly useful.

3.2. Preparation of requests stream representation

Requests stream representations can be prepared from several sources:

- server's logs;
- browser's logs;
- web cache's logs;
- traffic simulators;
- files prepared by hand.

In effect such representation contains URLs (uniform resource locators) of interesting documents as shown in Fig. 4.

It is important issue which source choose because the better representation of interesting request traffic we have the more applicable are the results of web cache configuration

[2] Software and documentation are available at http://www.squid-cache.org/

```
......
http://hel.cee.hw.ac.uk/Vortices/pages/QueryForm.html
http://www8.org/w8-papers/3a-search-query/semantic/semantic.html
http://sun3.ms.mff.cuni.cz/%7Edingle/webcoherence.html
http://www.arrowpoint.com/solutions/white_papers/renaissance.html
http://www.mnot.net/cache_docs/
http://www.eecs.harvard.edu/%7Evino/web/usenix.196/
http://www.surfnet.nl/innovatie/desire1/deliver/WP4/sumreport.html
http://www.imimic.com/WPfeatures.html
http://www.ariadne.ac.uk/issue21/web-cache/
http://mows.rz.uni-mannheim.de/mows/pub/paper/www6.html
http://www.scope.gmd.de/info/www6/technical/paper151/paper151.html
......
```

Fig. 4. Request stream representation sample.

analysis for our network. The natural choice is the third source: logs of the web cache which performance we want to improve. But there are situations when other sources are also valuable. For example if we try to predict some requests stream that can be specially troublesome for our network in the future we can use other sources where such traffic patters already occurred.

The use of each source has its advantages and disadvantages. Apart from requests simulating software the data obtained from presented sources needs some processing. This should be done very carefully since high quality of requests stream representation is crucial for further analysis.

Collecting requests data from web browser caches of individual users can be troublesome since such caches are unpredictably cleaned and their content cannot usually be obtained without permission.

Logs of WWW servers normally represent wider range of network users than browser caches that we can use. But this data is limited to only one place within network and access to such logs is often restricted.

On the contrary to those two sources web cache logs contains data with very adequate characteristic. Such stream contains requests from many users workstations directed to many servers with original content. But acquireing access to web cache logs with specially interesting requests patterns from outside our organization may also be difficult.

There are several problems with pre-processing log's data. For example often the result (stored in cache) of fulfilling a document request is not only the desired document but also some additional objects like accompanying graphics. Of course such objects does not belong to the original requests stream and should be removed from examined representation. Otherwise these objects are reported as cache hits and thus detoriate cache performance outcomes. There are some web sites that – when accessed – cause generation of new requests by browser (usually new browser windows appear) that are also stored in cache log. Very often these automatically generated requests refer to web pages that exist very shortly. In order to make results of web cache performance measurement repeatable such requests should be removed as well. Dynamically generated content of a web page also cause problems. Accessing such page results in different records for the same request for different times of access. This problem can be solved either by removing such requests from the request stream being prepared for examination or by proper counting different records for such request while computing performance measures.

Another important issue are documents specially marked to prevent them from caching because of the reasons presented in the previous chapter. Since such objects affect web cache performance they should be appropriately treated during the preparatory stage or while computing performance measures. It is worth to remove uncachable documents from the examined representation when speed of running experiments is important.

In order to address some of these problems a concept of two-phase measurement of web cache performance has been developed. This concept is more precisely presented in the next part of this chapter.

The disadvantage of simulated request stream is that such representation can be "too artificial" comparing to representations based on real traces, like computer logs. On the other hand simulation is more flexible. There is software[3] for simulating different request stream patterns. Such stream is generated according to given distribution of requests and number and characteristics of "virtual" clients and servers. The content of the servers are profiled. The space needed for requests storage is also reduced because instead of "real" some simplified request are used (they refer not to URLs but to some short identifiers). Thus processing such streams is also faster.

3.3. Proxy cache performance measurement

Metrics used for measuring web cache performance and factors influencing it have been introduced in the previous

[3] Web Polygraph software and documentation is available at http://polygraph.ircache.net

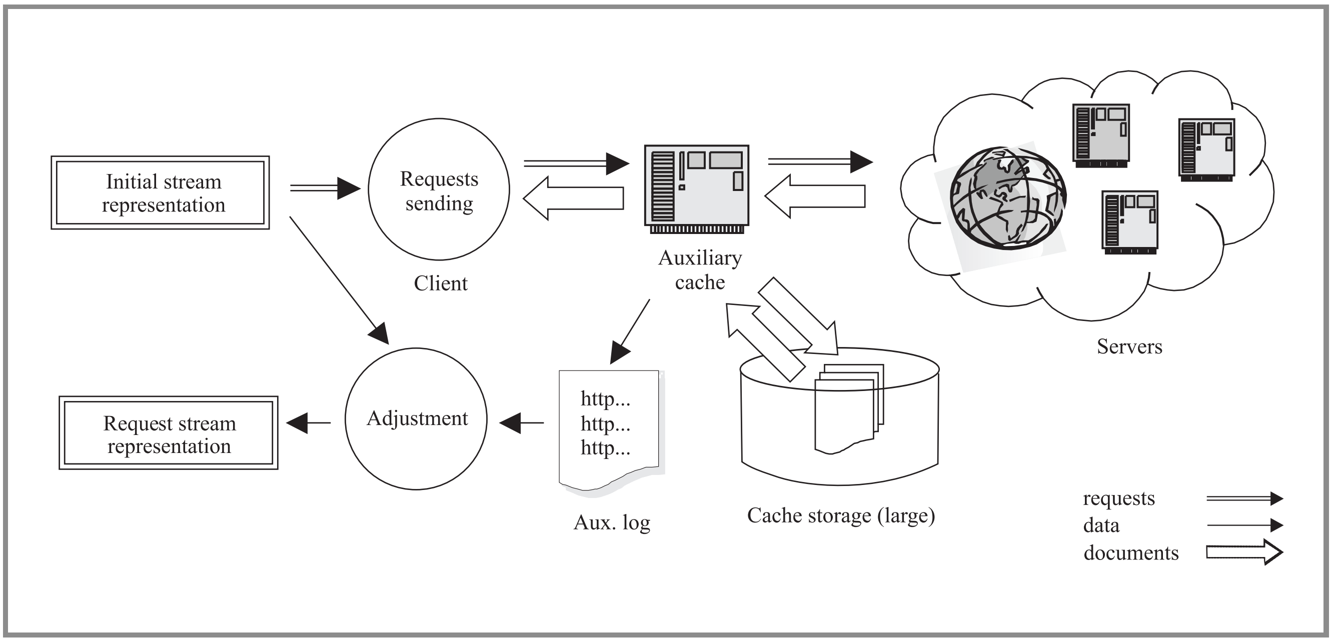


Fig. 5. Measurement phase 1 – preliminary data collection with use of auxiliary proxy cache.

chapter. The tool presented in this article has been designed to have capability to compute following measures:

- document hit ratio;
- byte hit ratio;
- average response time.

Computing the last metric has not yet been implemented. There is number of possible other metrics that can be derived from these basic ones.

Important issue connected with examining web cache performance is reducing the influence of unstability of external network – like Internet – for the results. This must be accomplished in order to make the experiments repeatable. It is crucial for valuable analyses to ensure that response times for the same requests are equal and web cache parameters values do not differ significantly. The design of experiments should also guarantee availability of responses for consecutive runs. In order to accomplish this aim web cache is examined in two phases using two joined web caches:

1) preliminary phase (connected with the second stage of the process);
2) examination phase.

3.3.1. Phase 1 – preliminary data collection

During this phase auxiliary web cache is used (Fig. 5). The requests from the examined stream representation are sent to this cache. In response auxiliary cache tries to retrieve requested documents from external network and deliver them to requester. Its storage space is large enough to store all the documents requested and its configuration is set up so that no documents are sweeped out because of their features like size. This phase is finished when all requests has been processed and auxiliary cache storage space is filled with all obtainable objects. To make results of the experiment more insensitive to external network failures this phase should be repeated when any undesirable network event occur. Log of the auxiliary cache contains values of response times for sent requests that can be further used for computing performance metrics. This phase is coupled with the data preparation stage. Requests stream representation is adjusted depending on the outcomes of this phase by removing some requests from the initial representation. Thus results of the next phase are not detoriated and the processing time is shorter.

3.3.2. Phase 2 – proxy cache performance examination

During this phase adjusted requests stream representation is used (Fig. 6). This time requests are sent to the actual web cache assigned for examination. This cache is connected to auxiliary cache so that requests sent to examined cache are fulfilled by retrieving documents from the auxiliary cache. Refering to external network is not needed.

In the course of experiment number of configuration alternatives are tested, accordingly to the plan saved in a special file introduced before. In each iteration one alternative is examined and performance data is stored in another log of actual web cache. Examining cache performance requires

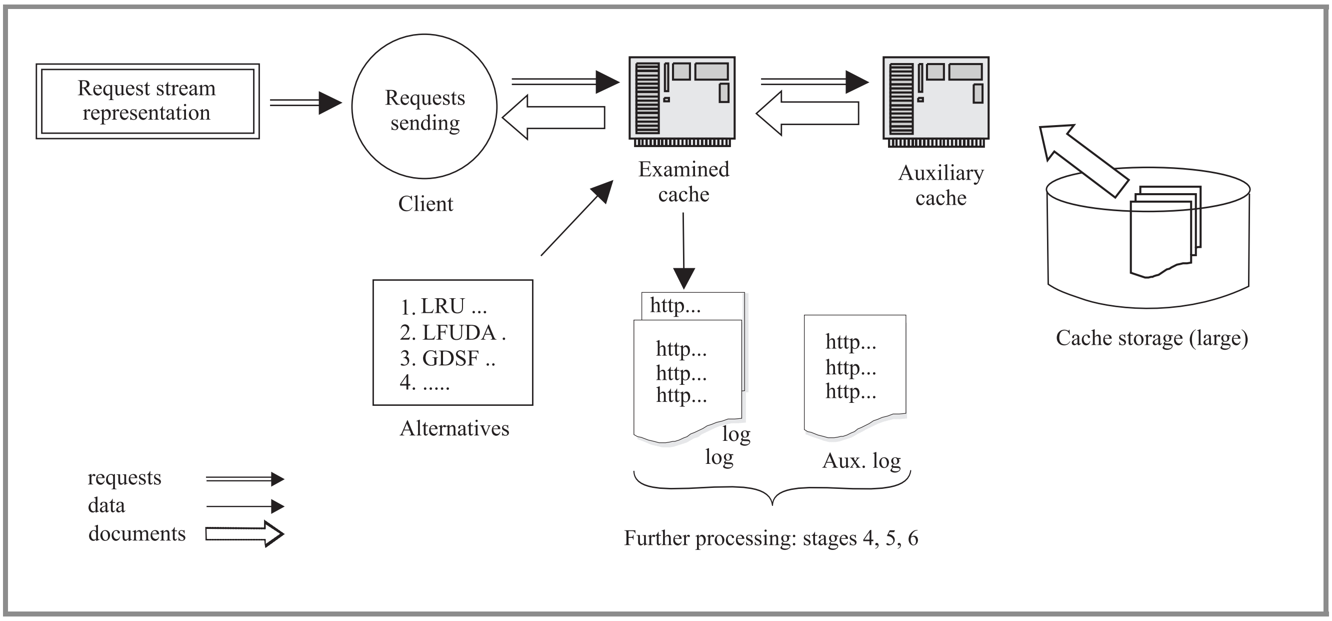


Fig. 6. Measurement phase 2 – proxy cache performance examination.

several technical tasks (they are automated by UNIX shell scripts):

1) resetting examined web cache to some initial state;

2) retrieving next set of parameters and setting them for examined web cache;

3) starting program sending request from examined representation to web cache;

4) log recycling.

3.4. Extraction of output variables

During this step desired performance measures are computed using the output data extracted from the examined and auxiliary web cache logs. This metrics are saved in a file where one row represents one configuration alternative and values of measures are kept in columns.

For the presented research purposes only three metrics were considered but number of other are possible to define depending on analytical needs. Thus number of criteria applied during multicriteria model analysis can be increased by computing new metrics here. In order to fulfill this task UNIX shell script has been created and used.

3.5. Generation of mathematical model

The problem of choosing the best one from a set of discrete alternative can be represented by the model shown in Table 1.

Because during the multicriteria analysis step criteria are chosen from the set of output variables condition $i <= m$, must be satisfied, where i is the number of criteria. Following equations complete formulating of the mathematical model for this problem is shown in Table 2.

Table 1
Discrete alternatives choice model

Alternatives	Output variables			
	y_1	y_2	$\dots$	y_m
z_1	a_{11}	a_{12}	$\dots$	a_{1m}
z_2	a_{21}	a_{22}	$\dots$	a_{2m}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
z_n	a_{n1}	a_{n2}	$\dots$	a_{nm}

Explanations: y – represents output variable (measure, metric); z – represents decision alternative (one configuration); m – is number of output variables; n – is number of decision alternatives; $a_{11}, \dots, a_{nm}$ – are output variables values.

This mathematical model is the basis for generating "core" model – some representation of regarded problem expressed in a mathematical programming languge. Such model comprises all physical and logical relations between variables describing the problem. The core model defines implicitly a set of feasible solutions but it does not include information about decision maker preferences.

The core model generator has been developed as a computer program written in C++ language[4]. This program has been tested for use on Sun Solaris platform. The output of this program is a file in LP-DIT format used by the MOMIP solver.

[4] This software has been created with cooperation with dr Marek Makowski.

Table 2

Bounds	Initial bounds	Criteria
$z_1+z_2+\ldots+z_n=1$	$y_1=a_{11}z_1+a_{21}z_2+\ldots+a_{n1}z_n$	$y_1=a_{11}z_1+a_{21}z_2+\ldots+a_{n1}z_n$
$z_j \ni \{0,1\}$	$y_2=a_{12}z_1+a_{22}z_2+\ldots+a_{n2}z_n$	$y_2=a_{12}z_1+a_{22}z_2+\ldots+a_{n2}z_n$
$j=1,\ldots,n$	....	....
$n=m+i$	$y_m=a_{1m}z_1+a_{2m}z_2+\ldots+a_{nm}z_n$	$y_i=a_{1i}z_1+a_{2i}z_2+\ldots+a_{ni}z_n$

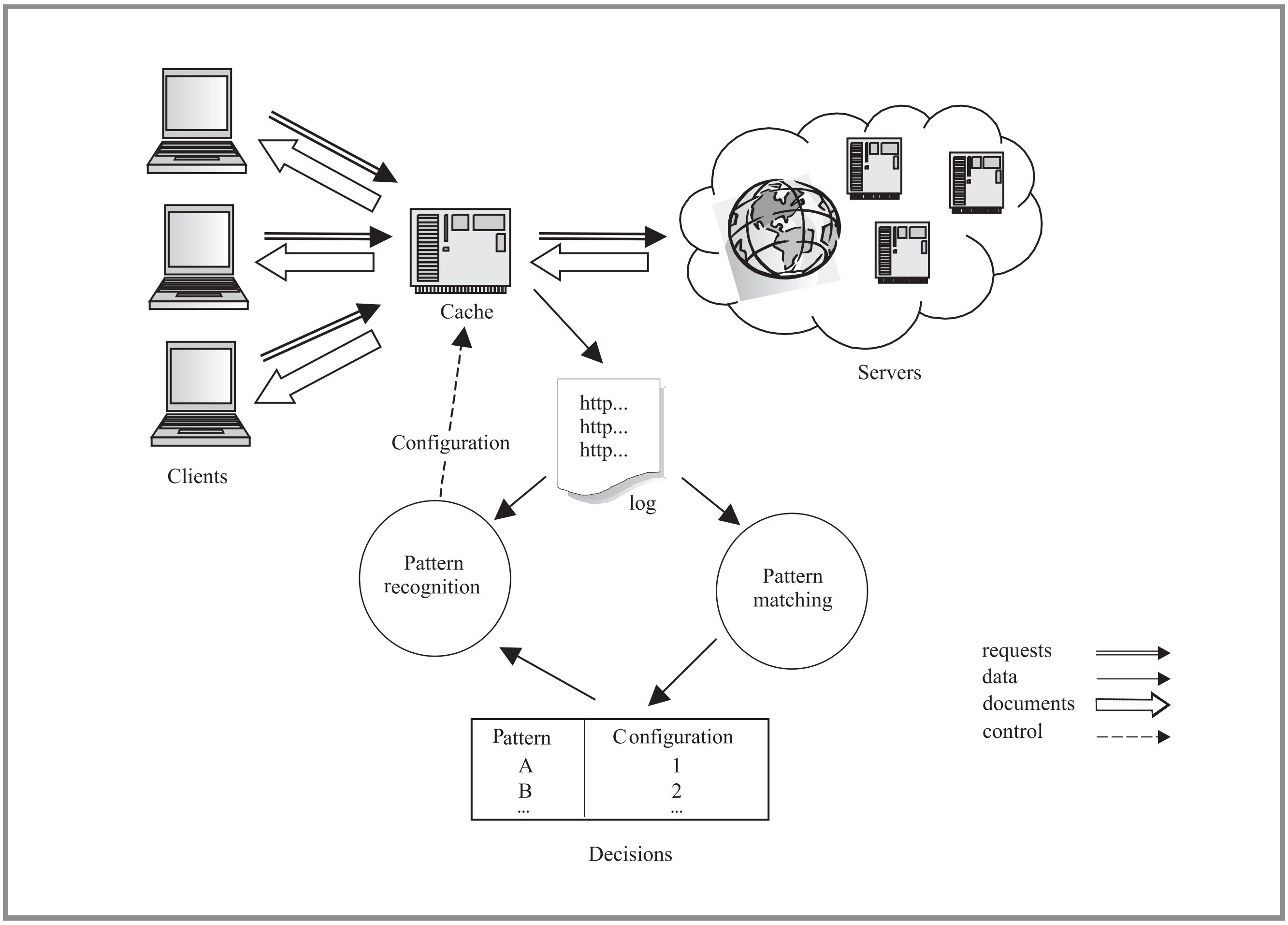


Fig. 7. Outline of the automated system for proxy cache configuration.

3.6. Multicriteria analysis of decision problem

The last stage of the process of analysis of decision alternatives for proxy cache configuration is multicriteria model analysis. For this purpose ISAAP modular tool is used. This software implements aspiration-reservation based decision support (ARBDS) approach. This methodology proved to be very successful and has found applications in many areas [2]. It is based on interactive analysis of a set of efficient solutions[5] of mathematical model representing multicriteria problem. Each iteration of the analysis is composed of specification of user preferences with regard to interesting criteria and computing corresponding solution for the problem. This procedure lasts until the most satisfying result for decision maker is found.

Since evaluating web cache performance usually demands several criteria this methodology seems very appropriate for this purpose. Results of simple experiment presented in the following chapter appear to confirm this choice.

ARBDS method can be summarized as a three-stage approach:

- specification and generation of a core model;
- preparatory stage;
- interactive procedure of analyzing efficient solutions.

The organization of process of analysis of web cache configuration presented in this paper is that the first stage of

[5] For efficient solutions no criterion can be improved without degrading a value of at least one other criterion.

ARBDS is handled by the program presented in the previous part of this chapter and the other two are managed by the ISAAP tool. Profound description of ISAAP functionality is beyond the scope of this paper and can be found in [2].

During the preparatory step of the method decision maker choses from the set of model variables criteria used for further problem analysis. For each criterion its type has to be specified:

- maximized – attaining maximum value for criterion is desired;
- minimized – attaining minimum value for criterion is desired;
- goal (stabilized) – achieving given goal (target) value of criterion is desired.

After some computations carried out by solver so-called **compromise solution** is found which corresponds to decision maker preferences set to outmost values (in case of MIP problem).

Aim of last step of ARBDS is to help decision maker find efficient solution corresponding the best to his/her preferences. During interactive procedure decision maker specifies his/her preferences and obtains another computed solution. Specification of preferences for each criterion is based on (see Fig. 12 for illustration):

- aspiration level – this value of criterion decision maker wants to achive;
- reservation level – this value of criterion decision maker wants to avoid.

This procedure is continued until decision maker is satisfied with solution or stops analysis.

3.7. Automated system for proxy cache configuration

In the future analytical environment for web cache management presented here can advance to automated system supporting proxy managers in improving configuration cache in response to changing requests stream patterns. Such system should incorporate, apart from the components presented above, also some software responsible for recognizing requests stream patterns and applying corresponding set of parameters stored in a repository to the managed proxy cache.

The role of the decision support tool should be updating repository of traffic patterns matched with cache configurations based on the analytical process described in this chapter. This idea is outlined in Fig. 7.

4. Simple experiment

Experiment presented below is very simple and any important conclusion concerning examined requests stream representation or web cache configuration could not be drawn. The aim of this experiment was to demonstrate application of presented approach to some not paper-based example.

4.1. Configuration alternatives

Three web cache parameters were used during this experiment (see chapter 2.3 for details):

- document replacement policy;
- **cache swap high** threshold;
- **cache swap low** threshold.

As shown in Fig. 3 three values of the first parameter were used during the test: `LRU`, `LFUDA` and `GDSF`. For second parameter also three values were applied: level of 80, 85 and 90 percent. Two values of the last used parameter were 95 and 100 percent. Altogether it makes up 18 combinations-alternatives of web cache configuration.

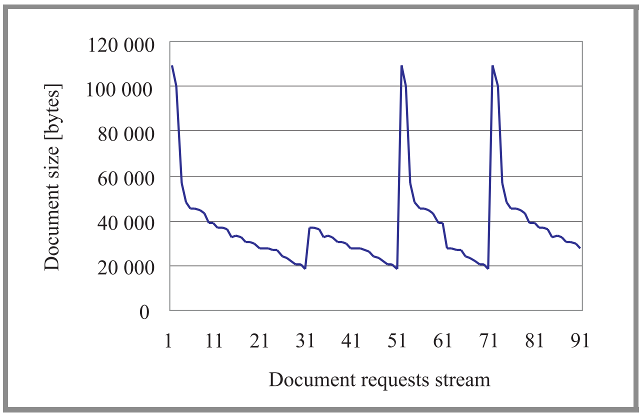


Fig. 8. Illustration of examined requests stream.

4.2. Requests stream representation

Representaion used in the experiment was prepared manually. Figure 8 shows characteristic of examined requests stream representation. This representation contains 100 requests for html documents concerning web cache issues. Document sizes vary from 20 to over 100 kilobytes. These documents were carefully selected from the larger set in order not to detoriate web cache performance outcomes. Only 30 of them were unique.

4.3. Experiment run

Firstly, prepared representation has been sent using WGET software to auxiliary web cache. Since requests for troublesome documents has been removed from the representation beforehand and auxiliary cache storage space was sufficiently large all the documents were stored in this cache. This operations were performed manually.

Secondly, web cache assigned for examination was connected to the auxiliary cache. UNIX shell script was run

Table 3
Performance results – efficient solutions are bolded

Alternative number	Parameters			**Measures**	
	replacement policy	swap lower threshold	swap upper threshold	**document hit ratio [%]**	**byte hit ratio [%]**
1	LRU	80	95	47	43
2	LFUDA	80	95	48	46
3	GDSF	80	95	46	35
4	LRU	85	95	46	43
5	LFUDA	85	95	44	36
6	GDSF	85	95	46	35
7	**LRU**	**90**	**95**	**50**	**46**
8	LFUDA	90	95	41	34
9	**GDSF**	**90**	**95**	**53**	**43**
10	LRU	80	100	48	46
11	LFUDA	80	100	30	34
12	GDSF	80	100	46	35
13	LRU	85	100	47	44
14	LFUDA	85	100	47	43
15	GDSF	85	100	46	35
16	LRU	90	100	46	42
17	LFUDA	90	100	44	37
18	GDSF	90	100	52	42

that automated tasks presented in the previous chapter. Requests from prepared stream representation were fed into examined cache and performance data was collected in log for each configuration alternative accordingly to the sequence stored in experiment plan file.

4.4. Performance results

After required data has been collected in cache logs values of two metrics were computed for each configuration alternative: document hit ratio and byte hit ratio. Results are shown in Table 3. Efficient solutions (alternatives 7 and 9) are bolded and further analysis concentrates on them. The table shows that for the given representation best results were achieved when applied document replacement politics was either LRU or GDSF. In both cases the values of upper and lower threshold for document sweeping were the same. These were also default values.

Figure 9 shows decision alternatives presented in criteria space. Number of points is less than number of alternatives since some alternatives correspond to same performance results. Black circles represent configuration alternatives for which solutions are not efficient. Efficient solutions are shown as rhombuses and these are most interesting to decision maker.

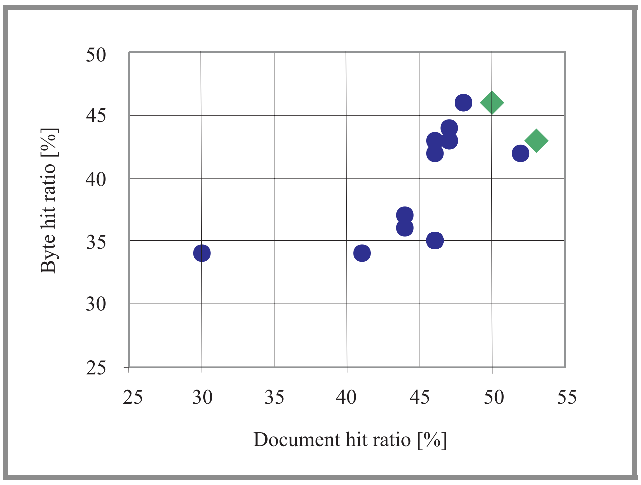
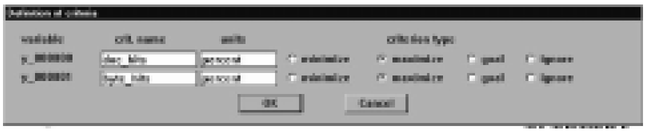


Fig. 9. Decision alternatives presented in criteria space.

4.5. Multicriteria analysis

Model generated using the software presented in previous chapter contained two metrics and 18 alternatives. These metrics were used as criteria in multicriteria analysis stage. It is natural to demand that document hit ratio and byte hit ratio were maximized. Therefore this type was selected for both criteria as shown in Fig. 10. Criteria names were chosen as doc_hits and byte_hits correspondingly.

First solution found was connected with alternative number 7. Corresponding criterion values are represented as white circles in Fig. 11. Decision maker preferences are set to criteria outmost values.


Definition of criteria

variable	crit. name	units	criterion type			
y_000000	doc_hits	percent	○ minimize	● maximize	○ goal	○ ignore
y_000001	byte_hits	percent	○ minimize	● maximize	○ goal	○ ignore

OK Cancel

Fig. 10. Criteria type specification.

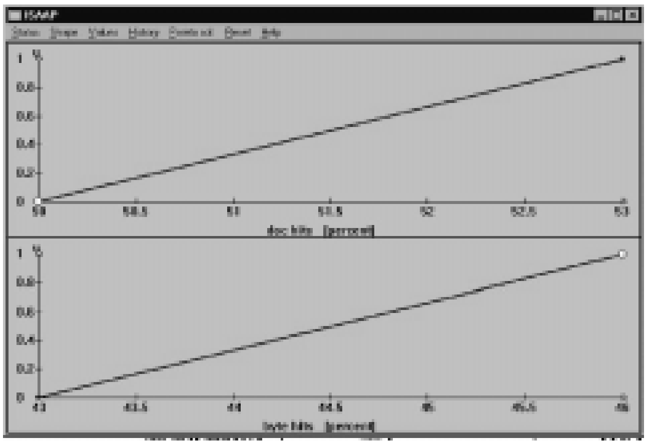


Fig. 11. Neutral solution (preferences set to criteria outmost values).

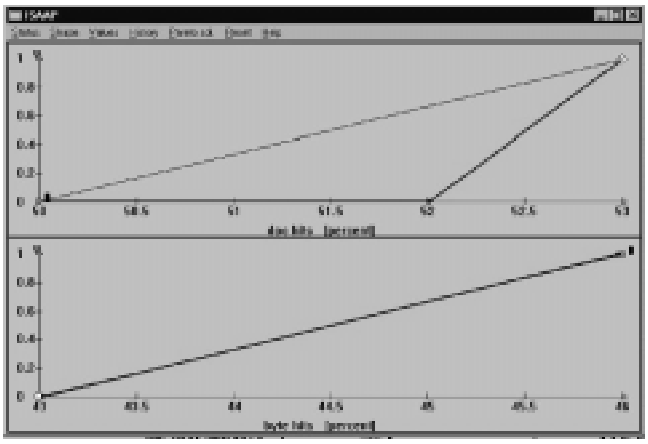


Fig. 12. Another solution – after preferences for the first criterion has changed.

In the next step preferences corresponding to doc_hits criterion were changed. Reservation level was set to 52%. Another efficient solution was computed. This time solution pertained to configuration alternative number 9. Value of the first criterion improved but at the expense of the value of second criterion (Fig. 12).

Because considered problem was much simplified there was no use to continue multicriteria analysis further. In this particular case it is hard to find justification for choosing one of these alternatives. Nevertheless in real situation when not miniature but significant requests stream representation and more criteria are involved, such analysis after several iterations may bring important results. Not only in terms of improved performance of web cache but also it could lead to better understanding by decision maker specific web cache behaviour and requests stream characteristic.

5. Summary

It has been shown that analysis of web cache performance can be presented as an analytical process. During this process several tasks have to be accomplished in order to transform input data to ready-to-analysis model. The concept introduced in this paper shows that this process can be backed by decision support tool helping in tuning web cache parameters.

Although results reported here are preliminary it seems that applying multicriteria model analysis in the field of web cache management can be fruitful. In order to advance presented tool to more mature status consultation of web cache expert is needed. Especially preparing requests stream representation and multicriteria analysis stages demand experienced user to achieve desired quality of such representation and to correctly interpret results of analysis.

The tool introduced here can be useful for persons dealing with web cache management in several ways. Firstly, it enables analysis of many configuration alternatives while number of interesting criteria are taken into account. Secondly, it can be used as a tool for testing cache functionality. This area is under fast development and new parameters are continously implemented. Their influence on web cache performance deserves scrutiny in specific network environment. Thirdly, when integrated with some additional software, it can be applied as an automated system for proxy cache configuration presented in chapter 3.

Acknowledgements

Work presented here was initiated during 3-month Young Scientists Summer Program organized by International Institute of Applied System Analysis in Laxenburg (Austria) in the year 2000. Author would like to thank dr Marek Makowski, dr Janusz Granat and Jerzy Sobczyk for their invaluable cooperation.

References

[1] J. Dilley, R. Friedrich, T. Jin, and J. Rolia, "Web server performance measuerement and modeling techniques", *Perform. Eval.*, vol. 33, pp. 5–26, 1998.

[2] J. Granat and M. Makowski, "ISAAP – interactive specification and analysis of preferences", *Interim Report*, IR-98-052, Nov., 1998.

Jarosław Pietrzykowski
e-mail: J.Pietrzykowski
National Institute of Telecommunications
Szachowa st 1, 04-894 Warsaw, Poland

Paper

Applications of commercial data analysis software for testing a telecommunications network

Jerzy Paczocha

Abstract — This paper shows practical aspect in application commercial software to analysis data concerning information from telecommunication network. The following items are discussed: example of network estimation methodology, the main advantages of network testing system delivering information concerning quality of service (QoS) and network performance (NP) from the user's perspective, general testing system description and data analysis model and example forms of presentation.

Keywords — network analysis, network estimation, network performance, quality of service, PSTN/ISDN.

1. Introduction

Commercial analysis software (computer decision support) are usually dedicated to business operations. Many functions of this software can be useful for processing data from telecommunication equipment (user traffic) and testing devices (testing traffic).

Application is presented as an example PSTN/ISDN network testing system for a telecommunication operator. It is based on experience from projects (AWP-IŁ) prepared by National Institute of Telecommunications in Warsaw for Polish Telecom (TP SA) and Polish Telecom Regulator (URT). First version uses relational data base and special made analysis software.

Implementation of computer decision software was necessary, because large volume data are processed. Short time calculation and many presentation forms are obtained.

During the choice of analysis software there were considering the following aspects:

- user friendly interface for analyst and common users;
- easy and flexibility creating new report forms of analysis and drilling if it is required;
- modelling form results without programming;
- possibility of next evolution forms presentation and stages analysis;
- time of calculation;
- easy distribution and security access to information;
- export report form to other applications;
- warranty of the next development and support.

2. Network estimation testing system

For general estimation quality of service, network performance and diagnostics may use statistical analysis calculating data, which are collected from, on example:

- network management centre;
- user traffic – selected tariffs records;
- signalling network;
- test traffic – the main subject of this presentation;
- subscriber port monitoring of exchanges;
- customer complaints – interview;
- faults reports.

Usually test traffic is generated no more than 2% of users traffic. Number and frequency of test calls must be equivalent for required quality accuracy and network diagnostics.

Knowledge of network configuration and principles traffic routing is required for drilling data.

Typical PSTN network has a hierarchical structure.

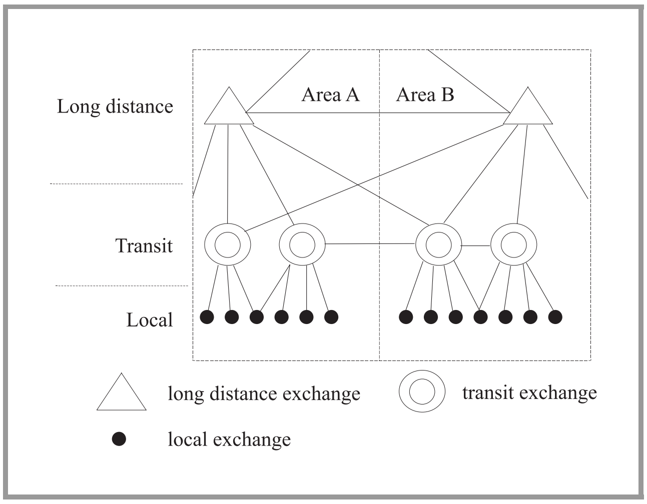


Fig. 1. Three levels of hierarchy of PSTN network configuration.

Figure 1 presents an example of network structure with three levels of hierarchy: local, transit and long distance. As a rule each exchange links minimum two trunk circuits to exchange in higher or the same level of network. Trunk circuit also connects transit exchanges.

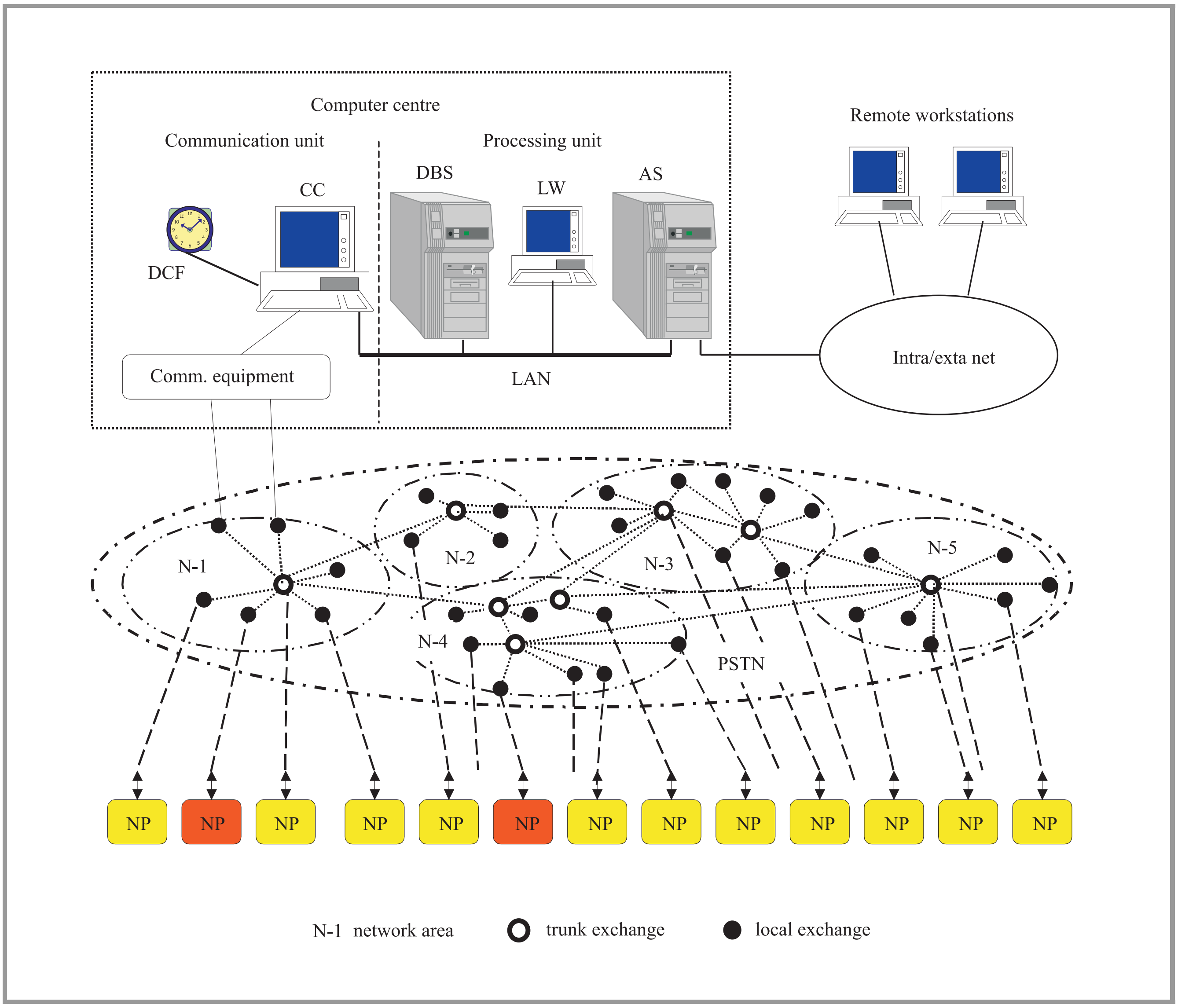


Fig. 2. Typical configuration of PSTN network test system.

Variants of connection string depend on traffic routing plan. Connection string is the result of call processing by exchange and it is depend on traffic load and failure. As a rule first route is selected to direct link, then to own trunk exchange and next to the other alternate routes.

System delivers information concerning the quality of service, network performance from the user's perspective. It indicates weak elements of the network.

System is useful for preparing the following estimation:

- quality of telephone services for calls: local, long distance and international;
- parameters: switching, transmission and billing integration;
- internetworking between the same sort of networks;
- internetworking between other sort of networks in the same service for example: voice and data;
- access to services for example: dialup Internet, information or audio services and other gateways.

System delivers comparative values independently from the sort of telecommunications equipment (exchanges) and personnel operation. It calculates statistics factors according to ITU-T and ETSI definitions.

These possibilities are helpful for telecoms in deregulated telecommunication market.

Typical testing system architecture consists of: network probes, computer centre and remote workstations (Fig. 2).

Computer centre is divided into processing and communication units.

Processing unit includes: data base server (DBS), analysis server (AS) and local workstations (LW).

Communication unit includes: communication computer (CC) and communication equipment for programming and transfer data with network probes.

Network parameters are measured by autonomous network probes (NP), which make test calls to each other. Network probes are connected to the network like normal subscriber. Network probes are controlled by computer centre.
Typical complete testing works includes following procedures:

- test planing – characteristics of testing traffic, number of test call and directions;
- generating control records for each test call;
- remote programming network probes;
- testing network by network probes;
- data collection;
- storage measuring record in data base;
- processing large data volume;
- presentation of results;
- distribution and control of the access to information.

Other useful functions are:

- reports from system configuration;
- access to the source measures record;
- security – distribution and qualify access to information and analysis according to the staff position in a telecommunication company;
- WEB user interface.

Data model includes definition and links for the following tables (for example):

1. Test description: test session, set of data probes which is used for test and time of test.
2. Data records from network probes: normalised data records are independent on coding data in probes and make it possible to compare to different telecommunication services.
3. Estimations of measuring records: values of data records are qualified to results in tree stage for drill.
4. Definition of calls: local, long distance and international.
5. Network probes configuration: port type, port number, place of installation and installation time.
6. Network test ports parameters: test port number, signalling parameters and electrical interface.
7. Network architecture description: network areas, hierarchy and telephone exchange.
8. Telecomm operator organization.
9. Country administrative organization.
10. Calendar: days of week, working days and holidays.

3. Examples of analysis

Interesting analysis presents behaviour of telecommunication network in time. Service accessibility performance parameter such as dial tone delay chart (Fig. 3) shows influence on the traffic intensity. It was chosen old electromechanical telephone exchange type PC 1000. It is sensitive to traffic load what is caused by equipment resources.

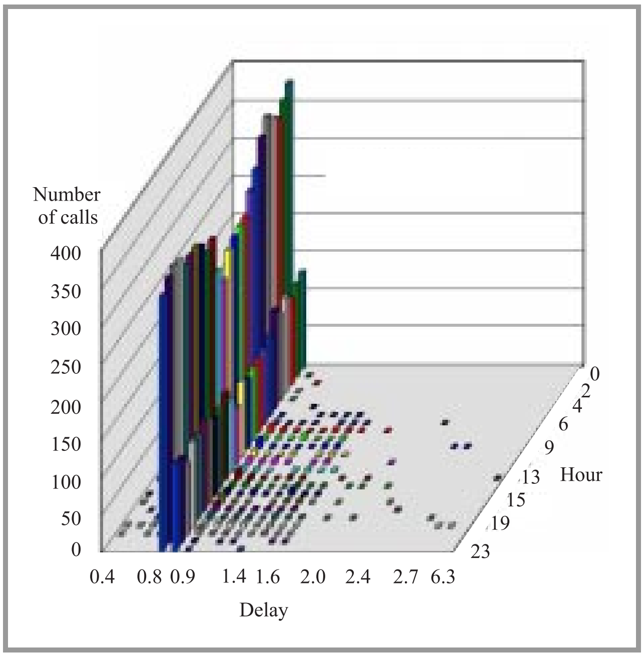


Fig. 3. Dial tone delay in hours for a PC 1000 telephone exchange.

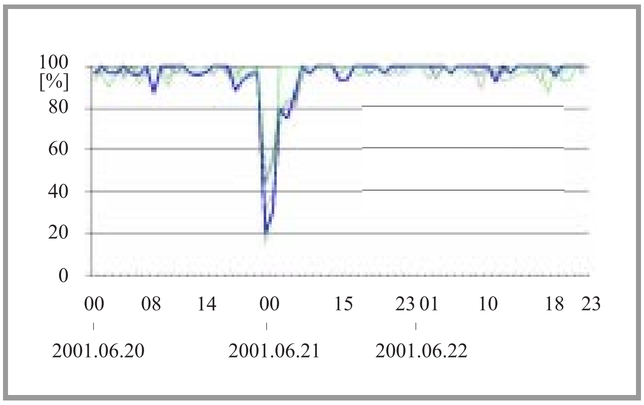


Fig. 4. Successful call ratio – relation from network area to the outside area.

During a peak of traffic at noon the share of lower values of dial tone delays goes down and higher delays go up. This situation is opposite at midnight. Usually peak of traffic is 10 times higher then average value of day and 10 times less at midnight.
Next chart (Fig. 4) shows the result of analysis of a part of PSTN network, which includes one transit and four sub-

ordinate telephone exchanges. The test calls were attempted from subscriber's ports of these exchanges to the other exchanges.

Because usually traffic from subordinate telephone exchanges is directed via superior transit exchange, successful call ratio parameter is related to superior transit exchange.This chart shows collapse of network about 2001.06.21 00.00 hour. This is the example how to apply this system to the network diagnostics.

4. Conclusion

Commercial software dedicated for business is useful only for simple statistics analysis for telecommunication.

It is possible to create reports in different configurations.

Analysis of telecommunications network, including hierarchy and change in time, needs data mining function with correlation.

References

[1] *Handbook on Quality of Service and Network Performance.* Geneva: CCITT, 1993.

[2] *Próbnik PM 3 systemu AWP-IŁ, przeznaczony do badania sprawności technicznej sieci PSTN/ISDN.* Dokumentacja techniczno-ruchowa. Warszawa: Instytut Łączności, 2001.

[3] *Travis Russell Telecommunications Protocols.* McGraw-Hill, 1997.

[4] "WALTER HAAS, TIQUS checks out quality and performance quality analysis in the telephone network", *Components*, no. 2, 1999.

[5] K. M. Brzeziński, *Istota sieci ISDN.* Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 1999.

Jerzy Paczocha was born on 15 January 1995 in Łódź. Graduated as Master of Engineering at the Faculty of Electronics, Warsaw University of Technology (WUT) in 1981. The main area he has dealt with are data communication, packed data network, remote control, telecommunication marketing, specialised telecommunication systems for the government administration, telecommunication traffic engineering and telecommunication network validation for the following companies such as Railway Scientific and Technical Centre, PTT Centre of Computer Technics, Banking Telecommunication Company, Central Board of Customs. Since 1995 he has been working in the National Institute of Telecommunications, Warsaw, Poland in "Research & Manufacturing Center for Information Technology and Services in Telecommunications" in field of the project and implementation of specialised computer systems for telecom operators and consulting services.

e-mail: J.Paczocha@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

INFORMATION FOR AUTHORS

The *Journal of Telecommunications and Information Technology* is published quarterly. It comprises original contributions, both regular papers and letters, dealing with a broad range of topics related to telecommunications and information technology. Items included in the journal report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

The *Journal is* dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, speech devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology. Suitable research-related manuscripts should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published by invitation only.

Papers published by invitation and regular papers should contain up to 15 and 8 printed pages respectively (one printed page corresponds approximately to 3 double-space pages of manuscript, where one page contains approximately 2000 characters).

Manuscript: An original and two copies of the manuscript must be submitted, each completed with all illustrations and tables attached at the end of the papers. Tables and figures have to be numbered consecutively with Arabic numerals. The manuscript must include an abstract limited to approximately 100 words. The abstract should contain four points: statement of the problem, assumptions and methodology, results and conclusion, or discussion, of the importance of the results. The manuscript should be double-spaced on only one side of each A4 sheet (210 × 297 mm). Computer notation such as Fortran, Matlab, Mathematica etc., for formulae, indices, etc., is not acceptable and will result in automatic rejection of the manuscript. The style of references, abbreviations, etc., should follow the standard IEEE format.

References should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. The **reference entry** (correctly punctuated according to the following rules and examples) **has to contain**.

From journals and other serial publications: initial(s) and second name(s) of the author(s), full title of publication (transliterated into Latin characters in case it is in Russian, possibly preceded by the title in Russian characters), appropriately abbreviated title of periodical, volume number, first and last page number, year. E.g.:

[1] Y. Namihira, „Relationship between nonlinear effective area and modefield diameter for dispersion shifted fibres", *Electron. Lett.*, vol. 30, no. 3, pp. 262-264, 1994.

From non-periodical, collective publications: as above, but after title – the name(s) of editor(s), title of volume and/or edition number, publisher(s) name(s) and place of edition, inclusive pages of article, year. E.g.:

[2] S. Demri, E. Orłowska, „Informational representability: Abstract models versus concrete models" in *Fuzzy Sets, Logics and Reasoning about Knowledge*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301-314.

From books: initial(s) and name(s) of the author(s), place of edition, title, publisher(s), year. E.g.:

[3] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

Figure captions should be started on separate sheet of papers and must be double-spaced.

Illustration: Original illustrations should be submitted. All line drawings should be prepared on white drawing paper in black India ink. Drawings in Corel Draw and Postscript formats are preferred. Colour illustrations are accepted only in exceptional circumstances. Lettering should be large enough to be readily legible when drawing is reduced to two- or one-column width – as much as 4:1 reduction from the original. Photographs should be used sparingly. All photographs must be gloss prints. All materials, including drawings and photographs, should be no larger than 175 × 260 mm.

Page number: Number all pages, including tables and illustrations (which should be grouped at the end), in a single series, with no omitted numbers.

Electronic form: A floppy disk together with the hard copy of the manuscript should be submitted. It is important to ensure that the diskette version and the printed version are identical. The diskette should be labelled with the following information: a) the operating system and word-processing software used, b) in case of UNIX media, the method of extraction (i.e. tar) applied, c) file name(s) related to manuscript. The diskette should be properly packed in order to avoid possible damage during transit.

Among various acceptable word processor formats, $\mathrm{T_EX}$ and $\mathrm{LAT_EX}$ are preferable. The *Journal's* style file is available to authors.

Galley proofs: Proofs should be returned by authors as soon as possible. In other cases, the article will be proofread against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within two weeks after receiving the offprints.

The *copy of the „Journal"* shall be provided to each author of papers.

Copyright: Manuscript submitted to this journal may not have been published and will not be simultaneously submitted or published elsewhere. Submitting a manuscript, the authors agree to automatically transfer the copyright for their article to the publisher if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and also all translation rights. No part of the present journal may be reproduced in any form nor transmitted or translated into a machine language without permission in written form from the publisher.

Biographies and photographs of authors are printed with each paper. Send a brief professional biography not exceeding 100 words and a gloss photo of each author with the manuscript.

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Editorial Office

tel. +48(22) 872 43 88
tel./fax:+48(22) 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.itl.waw.pl/jtit