JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

This issue of the *Journal of Telecommunication and Information Technology* continues topics addressed in the previous issues related to multiple criteria decision making, data mining, knowledge acquisition, management and engineering, and other advanced information technologies with application to telecommunications and other network services, but starts with two more fundamental papers, one concerning philosophy of technology in time of informational revolution and the second, invited paper on methodology of multiple criteria decision making. This is continued with papers on applications of multicriteria analysis to telecommunications and on knowledge management in telecommunications research institute. This issue contains also papers on diverse technological and economic aspects of telecommunication technology, such as spectral division of the optical fiber passband, code cardinality for ultra wideband systems, high-frequency power amplitude modulators, cooperative and non-cooperative, integrative and distributive market games, problems of broadband in rural areas, etc.

The first paper, by Andrzej P. Wierzbicki from National Institute of Telecommunications in Warsaw, *Delays in Technology Development: Their Impact on the Issues of Determinism, Autonomy and Controllability of Technology*, provides a discussion of diverse delays occurring in technology development, and an explanation of reasons why, when seen holistically from outside, the process of technology development might appear as an autonomous, self-determining, uncontrollable process. When seen from inside, however, the process is far from being uncontrollable. This paradox is explained by the fact that technology development contains many processes with delays, in total amounting sometimes to fifty years; when seen from outside, such a process might appear uncontrollable, even if it is very much controllable when approached internally and in detail. Therefore, the definition and types of technology creation as well as stages of technological processes are discussed in some detail in this paper. Some aspects of the contemporary informational revolution and some recent results on micro-theories of knowledge and technology creation are also reviewed. It is suggested that one of possible ways of changing the paradigmatic attitude of philosophy of technology is to invite some such philosophers to participate in the development of modern tools of knowledge civilization era, such as software development and evaluation. The conclusions of the paper stress the need of essentially new approaches to many issues in the time of informational revolution.

The second, invited paper by Włodzimierz Ogryczak from Warsaw University of Technology, *Reference Point Method with Importance Weighted Partial Achievements*, stresses that the reference point method (RPM) is a convenient technique for interactive analysis of the multiple criteria optimization problems. It provides the decision maker (DM) with a tool for an open analysis of the efficient frontier. The interactive analysis is navigated with the commonly accepted control parameters expressing reference levels for the individual objective functions. The partial achievement functions quantify the DM satisfaction from the individual outcomes with respect to the given reference levels. The final scalarizing achievement function is built as the augmented max-min aggregation of partial achievements with respect to the given reference levels which means that the worst individual achievement is essentially maximized but the optimization process is additionally regularized with the term representing the average achievement. In order to avoid inconsistencies caused by the regularization, the max-min solution may be regularized by the ordered weighted averages (OWA) with monotonic weights which combines all the partial achievements allocating the largest weight to the worst achievement, the second largest weight to the second worst achievement, and so on. Further following the concept of the weighted OWA (WOWA) the importance weighting of several achievements may be incorporated into the RPM. Such a WOWA RPM approach uses importance weights to affect achievement importance by rescaling accordingly its measure within the distribution of achievements rather than by straightforward rescaling of achievement values. The recent progress in optimization methods for ordered averages allows one to implement the WOWA RPM quite effectively as extension of the original constraints and criteria with simple linear inequalities.

The third paper, by Michał Majdan from Warsaw University of Technology and National Institute of Telecommunications in Warsaw, *On Subjective Trust Management*, stresses that trust and reputation management is gaining nowadays more attention then ever as online commodity exchange and other open virtual societies became a widespread reality. Most widely used computational models use reputation metrics as global property assigned to each party. More sophisticated models try to use reputation as subjective property. While introducing subjective reputation there arise a need to model preferences of agents. The paper proposes to use weighted ordered weighted average operator to support the decision maker in assessing available evidence about other's party behavior.

The fourth paper, by Paweł Białoń from National Institute of Telecommunications in Warsaw, *An Eclectic Approach to Network Service Failure Detection Based on Multicriteria Analysis with an Example of Mixing Probabilistic Context Free Grammar Models*, presents a method of failure detection in telecommunication networks. This is a meta-method that correlates alarms raised by failure-detection modules based on various philosophies. The correlation takes into account two main characteristics of each module and the whole meta-method: the percentage of false alarms and the percentage of omitted failures. The trade-off between them is tackled with aspiration-based multicriteria analysis. The alarms are correlated using linear classification by support vector machines. An example of the profitability of correlating alarms in such way is shown.

The fifth paper, by Joana Fialho, Pedro Godinho, João Paulo Costa, Ricardo Afonso, and José Gonçalo Regalado from University of Coimbra and PT Inovação in Aveiro, Portugal, *A Level-Based Approach to Prioritize Telecommunications R&D*, presents an approach to evaluate R&D projects in telecommunications. These projects have particular features that cannot be properly incorporated by classical evaluation methods. This approach incorporates different criteria, both quantitative and qualitative, and also management flexibility and uncertainty. Thus, it is an approach that can be applied to real data of R&D projects in a telecommunications company.

The sixth paper, by Cezary Chudzian from National Institute of Telecommunications in Warsaw, *Ontology Creation Process in Knowledge Management Support System for a Research Institute*, documents a part of the research activities performed at National Institute of Telecommunications, related to development of research institute knowledge management support system. The ideas lying in the background of the system come from the recent theories of knowledge creation and creativity support and from experience with everyday practice of knowledge management in market companies.

The seventh paper, by Jarosław Sobieszek from National Institute of Telecommunications in Warsaw, *Towards a Unified Architecture of Knowledge Management System for a Research Institute*, presents some elements of architecture of planned knowledge management system dedicated to research institutions. Main contributions include social extension of the idea of adaptive hermeneutic agent and preliminary implementation of domain specific language for development of knowledge management systems.

The eighth paper, by Kamil Kołtyś, Piotr Pałka, Eugeniusz Toczyłowski, and Izabela Żółtowska from Warsaw University of Technology, *Multicommodity Auction Model for Indivisible Network Resource Allocation*, presents a multicommodity auction model that allocates indivisible network resources among bidders. The approach can be considered as a generalization of the basic multicommodity model for balancing communication bandwidth trade (BCBT). The BCBT model assumes that offers concerning inter-node links and point-to-point bandwidth demands can be realized partially. However, in the real-world trade there might be a need to include capacity modularity in the market balancing process. Thus the paper presents a model for balancing communication bandwidth trade that takes into account the indivisibility of traded bandwidth modules. This requires to solve a mixed integer problem and increases computational complexity. Furthermore, the pricing issue appears nontrivial, as the dual prices cannot be longer used to set fair, competitive market prices. For clearing the market, we examine the multicommodity pricing mechanism based on differentiation of buy and sell market prices.

The ninth paper, by Igor Goncharenko, Alexander Esman, Grigory Zykov, Vladimir Kuleshov, Marian Marciniak, and Vladimir Pilipovich, from Institute for Command Engineers, Belarus, Institute of Physics, National Academy of Sciences, Belarus, and National Institute of Telecommunications in Warsaw, *Spectral Division of the Optical Fiber Passband Using Narrowband Controllable Filter on the Base of Semiconductor Waveguide Microresonator*, analyzes a new principle of multichannel spectral division of optical fiber passband using controllable narrowband integrated optical filters composed of two-coupled ring microresonators made of different semiconductor materials.

The tenth paper, by Mohammad Upal Mahfuz, Kazi M. Ahmed, and Nandana Rajatheva from University of Calgary, Canada, and Asian Institute of Technology, Thailand, *On the Effects of Code Cardinality for TH-PPM Ultra Wideband Systems*, demonstrates the effects of code cardinality at transmitter section on bit error rate (BER) performance of time hopping pulse position modulation (TH-PPM) based ultra wideband (UWB) indoor radio communication. In the transmitter, different code cardinality values have been chosen and correspondingly the effects on BER of the system have been investigated.

The eleventh paper, by Juliusz Modzelewski and Mirosław Mikołajewski from Warsaw University of Technology, *High-Frequency Power Amplitude Modulators with Class-E Tuned Amplifiers*, stresses that a high-frequency (HF) power amplifier used in a drain amplitude modulator should have linear dependence of output HF voltage V_o versus its supply voltage V_{DD}. This condition, essential for obtaining low-level envelope distortions, is met by a theoretical class-E amplifier with a linear shunt capacitance of the switch. The paper analyzes the influence of non-linear output capacitance of the transistor in the class-E amplifier on its $V_o(V_{DD})$ characteristic using simulations of the amplifiers operating at frequencies 0.5 MHz, 5 MHz, and 7 MHz.

The twelfth paper, by Sylwester Laskowski from National Institute of Telecommunications in Warsaw, *Cooperative and Non-cooperative, Integrative and Distributive Market Games with Antagonistic and Altruistic, Malicious and Kind Ways of Playing*, illustrates distinctions between important concepts of game theory, which support understanding the relation between subjects on competitive and regulated telecommunications services market. Especially it shows that often used distinction between retail and wholesale market that treat them respectively as competitive and cooperative can be misleading or even wrong.

The thirteenth paper, by Paweł Białoń from National Institute of Telecommunications in Warsaw, *Problems of Broadband in Rural Areas in Light of the BReATH Project Experiences*, presents some lessons learned from the EU project "Broadband e-Services and Access to the Home" (2005–2007) concerning the broadband development in rural areas. In particular, the paper discusses the common problems of broadband deployment in the rural environment, various aspects of stimulating demand for broadband, the limitations of public aid and, most importantly, the problems of techno-economic analysis.

Andrzej P. Wierzbicki
Guest Editor

Paper

Delays in Technology Development: Their Impact on the Issues of Determinism, Autonomy and Controllability of Technology

Andrzej P. Wierzbicki

Abstract—The paper provides a discussion of diverse delays occurring in technology development, and an explanation of reasons why, when seen holistically from outside, the process of technology development might appear as an autonomous, self-determining, uncontrollable process. When seen from inside, however, the process is far from being uncontrollable. This paradox is explained by the fact that technology development contains many processes with delays, in total amounting sometimes to fifty years; when seen from outside, such a process might appear uncontrollable, even if it is very much controllable when approached internally and in detail. Therefore, the definition and types of technology creation as well as stages of technological processes are discussed in some detail in this paper. Some aspects of the contemporary informational revolution and some recent results on micro-theories of knowledge and technology creation are also reviewed. It is suggested that one of possible ways of changing the paradigmatic attitude of philosophy of technology is to invite some such philosophers to participate in the development of modern tools of knowledge civilization era, such as software development and evaluation. The conclusions of the paper stress the need of essentially new approaches to many issues in the time of informational revolution.

Keywords— autonomy, change of episteme, delay time, determinism and controllability of technology, impacts of informational revolution, paradigm of philosophy of technology, technology development and evaluation.

1. Introduction

Seemingly and actually, software development and evaluation is very distant from philosophy of technology. Software development and evaluation is detailed, specific, motivated by the goal of producing best, reliable and user-friendly software, applies specific staged development and evaluation processes as well as software quality criteria; it requires deep specialized knowledge about software engineering, and is future-oriented, concentrates *ex ante* on new products. Philosophy of technology is general, sees technology as a socio-economic system of producing and utilizing products of technology; sometimes accuses this system of being autonomous or deterministic – that is, developing according to its inner momentum, without taking into account humanistic values; often accuses this system of being unethical – underestimating technological risks; is historically oriented, concentrates on *ex post* evaluation of results of technological development.

Yet this does not mean that the visible gap between software evaluation and philosophy of technology is justified; nor that it is desirable. Software development will (or already has) become the decisive factor in the development of technology; also, it contributes to technological risks. Without including aspects of philosophy of technology, software evaluation is liable to be accused of *technological, instrumental and functional rationality*[1]; more seriously speaking, inputs from philosophy of technology might enrich software development and evaluation. On the other hand, without participating in software development, particularly in software evaluation, in times of information revolution, philosophy of technology runs the risk of becoming outdated and sterile. The conclusion is that both sides might gain by bridging the gap. However, we shall see that the initiative must come from software engineering side, simply because philosophy of technology is too paradigmatic – and I am telling this both as a technologist, since fifty years specializing in computer simulation and diverse related aspects of information technology, and a specialist working in recent years close to philosophy, on the new micro-theories of knowledge and technology creation.

This too paradigmatic attitude of philosophy of technology can be best illustrated by the opinion of Val Dusek [1], a leading humanist philosopher of technology, who even today denies the concept of informational revolution and calls all the discussion of the change of civilization era, of postindustrial, postcapitalist, informational or networked society a *technocratic hype* and *technological determinism*. On the other hand, as shown, e.g., in [2], the evidence of tremendous social and economic changes already occurring due to the impact of computing and network technology is obvious. We might add here that the automation and robotization of manufacturing already resulted in advanced countries in an essential *dematerialization of work* which contributed to the de-legitimization of the Marxian concept of the leading role of proletariat and thus to the fall of communist system. Thus, positions denying the change observed today correspond to closing eyes when spotting unpleasant objects. It might be related to an intuitive, unpleasant perception that if the thesis about an informational

[1] I just quote here typical phrases of philosophy of technology, even if I disagree with their meaning and use – because personally I see *technology as the art of creating tools*, in a broad sense including software, and refuse to accept the reduction of creative technological rationality to instrumental and functional aspects.

revolution leading to a new era is valid, then the classical philosophy of technology does not have a chance: it must address quite new themes and must ask technologists about advice, while it succeeded until now to concentrate on the criticism of the old industrial society and develop practically without any feedback from engineers.

Thus, the motivation of this paper is to outline a list of new topics of philosophy of technology, important in the times of informational revolution and the beginnings of a new era. We should start, however, with a criticism of a myth of old philosophy of technology, concerning the assumed (arbitrarily and intuitively, thus deeper than paradigmatically – actually, in the hermeneutical horizon[2] of old philosophy of technology) autonomy and determinism of technology.

2. The Reasons of Seeing Technology as an Autonomous, Deterministic System

We should recall first that the old philosophy of technology understands its object, the concept of *technology*, in diverse meanings (often without specifying the meaning used in a given discourse), but most often *as the socio-economic system of creating and utilizing products of technology or technological artefacts approached holistically*, while technologists tend to understand their field more narrowly, as *the art of creating tools and technological artefacts*. However, we shall discuss these distinctions in more detail later, here we concentrate on the properties of the socio-economic technological system. This system was often seen by the old philosophy of technology as autonomous, i.e., uncontrollable in technical terms, and deterministic, in at least two senses: self-determining (which is similar to autonomous) or determining the development of society. The latter is an obvious error when seen by a technologist who knows well that *technology proposes and society chooses*, although historically we can list such technological developments (Johann Gutenberg, James Watt, personal computers and computer networks) that *enabled* great economic, social and cultural changes; thus, *technology does not determine*, *only enables social changes*. The issue of self-determination and autonomy or uncontrollability of technology is more complicated, however.

The socio-economic system of creating and utilizing products of technology is complex. By approaching it holistically, without analyzing in detail its parts and their relations, the impression that this system is autonomous and self-determining is very likely to emerge. The most important reason for that impression might be the fact – overlooked by most philosophy of technology – that *this system includes many delays*. By delay we understand the time interval between starting an activity and observing its' results; obviously, in the development of technology we can observe at least the delay between starting a design and finishing it, including initial testing and evaluation. However, this delay is relatively small when compared to other delays in the processes of social acceptance and market penetration of products of technology. At the very beginning, new technological ideas appear often in academic communities; the character of knowledge creation in these communities is different than in industrial research organizations, see [4], [5] and Section 4 of this paper; this makes difficult the transfer of ideas from academia to industry and induces additional delays.

Even if a product is ready for market penetration, consumers initially distrust new products; it needs time to develop social demand. Moreover, oligopolistic firms on high technology markets delay acceptance of new standards, trying to preserve this way their markets shares; this is another reason of delays. These diverse socio-economic reasons increase the total delay between an original idea and its broad socio-economic use. *In the case of mobile cell telephony this delay amounted to fifty years* (the principle was developed for military purposes during the Second World War in the forties, broad social use occurred in the nineties of the 20th century). In the case of transistors and integrated circuits the delay was shorter, because of their importance in the time of cold war; but in the case of digital television the delay again exceeds fifty years. For other examples of such delays, see [6].

Now, a system with delays, if approached holistically from outside, very likely appears as autonomous and self-determining; we seem to have lost control over its functioning. This is very well known to specialists in control of systems with delays[3], but might require a more detailed explanation for non-specialists. Delay is a concept from systems dynamics, better known to technological systems dynamics studying systems with both inertial and pure delays than to sociological systems dynamics that by delays understands mostly inertial delays. By inertial delay we mean delay occurring as a result of accumulation processes, such as filling a glass with water; you can try to control it by increasing the volume of the stream of water. By pure delay we mean delay due to transportation, such as the delay occurring when you wait on an airport at a luggage conveyor, say, the delay between your luggage appearing on the conveyor and its' coming to the place where you wait for it. If you cannot move towards your luggage, you obviously lose control over it until it comes to your place.

Thus, when approaching from outside a socio-economic system with delays coming to fifty years, you certainly perceive a loss of control over the system. But how to ef-

[2] By a *hermeneutical horizon*, as specified more precisely in [3] though used earlier in diverse writings of hermeneutical philosophy, we understand a*n intuitively assumed system of beliefs in the truth of basic axioms*. A hermeneutical horizon is usually not expressed explicitly, but can be reconstructed, i.e., inferred from diverse clues. A hermeneutical horizon is thus an intuitive, deep foundation of a paradigm.

[3] Such as myself: long ago, I have worked intensively on industrial control of processes with delays and published in 1970 a paper on the maximum principle (a necessary condition of optimality of dynamic control) for systems with non-trivial pure delays in control, see [7].

fectively control systems with delays? There are several ways known to specialists; all of them, however, reduce to trying to anticipate its behavior, or at least measure or acquire information about this behavior with less delay than in the end effect. In the example with airport luggage conveyor, this amounts to choosing such a place that you can observe your luggage from the moment of its appearance on the conveyor and react appropriately when it falls down from the belt, or is taken by mistake by another passenger. How to use this analogy for controlling the development of technology? We must simply abandon the holistic, outside approach, analyze the details of the development and see in which points, at what stages of the process we can obtain anticipating information.

Therefore, we must simply abandon the position originated by a classical author of philosophy of technology, Jacques Ellul [8], limiting his interests to collective processes in the society that must be approached holistically, and followed – for diverse reasons – by most philosophers of technology. For example, Carl Mitcham [9, p. 65] argues that humanist philosophers, dominating philosophy of technology[4], simply cannot learn the details of technology, because "becoming mired in the specialized details of technology and its many processes tends to obscure relationships to nontechnological aspects of the human". However, *the main point of this paper is that we have here a binary, either-or choice*: either philosophers of technology continue to abstain from going into details of the process of technology creation, thus they will continue to see technology as a dark, uncontrollable force; or they will try to cooperate in effective control of technology, but then they must learn details. Such learning of details starts with the definition of technology and stages of technological processes.

3. What Is Technology and Stages of Technology Creation and Utilization

We can start by asking the question: *what is technology*? There are diverse answers to this question. Technology might be:

- for a philosopher of technology: *the socio-economic system of creating and utilizing technology*;
- for a postmodern humanist scientist: *an autonomous force enslaving humanity*;
- for an economist: *a way of doing things, a technical process*;
- in common language: *a technical artefact*;
- for a natural scientist: *an application of scientific theories*;
- for a technologist: *the art of constructing tools, an inherent faculty of humanity, motivated by the joy of creation*:
  - *liberating* people from hard work;
  - *helping* technology brokers (venture capitalists, bankers, managers) to make money – and if any effect of that is *enslaving*, the brokers are responsible;
  - *stimulating* the development of hard science by *inventions* which give it new principles to develop new concepts.

If there are that many answers, this means that the word *technology* is commonly used imprecisely, such as in common language it often means *a technological artefact*, while I rather use the term *product of technology* to denote this meaning. Being a technologist, I believe that our, technological understanding is most close to the essence[5] of the meaning of the word technology; however, since others might contend this interpretation (and Dusek [1] does not even list it in his discussion of definitions of technology), I agree to designate it *technology proper*. Moreover, it is very close to one of interpretations of the word technology by Martin Heidegger [10] – even if he used several such interpretations, selecting a convenient interpretation for a given discourse – as well as to the classical Greek word *techne*. In [11] and [12] the following definition was proposed:

Technology proper is a basic human faculty that concentrates on the creation of tools and artefacts needed for humanity in dealing with nature. It presupposes some human intervention in nature, but can also serve the goal of limiting such intervention to the necessary scale. It is essentially a truth-revealing, creative activity, thus it is similar to arts. It is also, for the most part, a problem-solving activity, concentrating on solving practical problems.

Philosophy of technology often says that the old concept of techne was changed by modern mass production, but this is mixing technology proper with mass production technological processes that constitute another stage of the socio-economic system of technology creation and utilization. Techne, technology proper, remains essentially the same: a truth-revealing, creative activity of constructing tools – naturally, tools characteristic for a given civilization era; we can speak thus about *techne*$_1$ in ancient Greece, *techne*$_2$ in the times of constructing telescopes and mechanical clocks, *techne*$_3$ in the era of industrial civilization, *techne*$_4$ in times of informational revolution and knowledge civilization, when the main tools constructed are software tools.

Now we should outline shortly – see [5] for a more detailed discussion – the relations of technology proper to hard science (natural sciences and mathematics) and to soft science (social sciences and humanities), as well as to the system

[4] Philosophers of mathematics are almost all – with a few exceptions – mathematicians; philosophers of technology are almost all – with even fewer exceptions – humanists or sociologists, not technologists.

[5] With all reservations concerning the possibility or rather impossibility of reaching the true essence of meanings.

of socio-economic applications of technology. They are outlined by the second part of the definition:

Thus, technology proper uses the results of basic sciences, if they are available; if they are not, technology proposes its own solutions, often promoting in this way quite new concepts, which are assimilated after some delay by the hard or social sciences. It is not an autonomous force, because it depends on all other human activities and influences them in return. It is, however, sovereign, in the same sense as arts are sovereign human activities. Autonomous forces can be found in the socio-economic system of applications of technology proper.

How, then, do the hard, basic sciences and technology depend on each other? As in many questions of human development, they influence each other through a positive feedback loop, see Fig. 1; technological development stimulates basic science, while scientific theories are applied technologically.

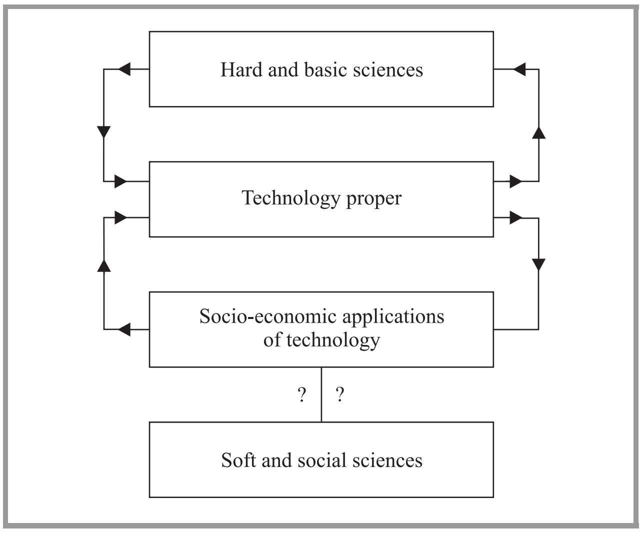


Fig. 1. Two positive feedback loops.

Recall that feedback – the circular impact of the time-stream of results of an action on its causes – was used by James Watt in a negative feedback loop and reinvented by Harold Black [13][6]. Feedback can be of two types: positive feedback when the results circularly support their causes, which results in fast development, like a growing avalanche, and negative feedback when the results circularly counteract their causes, which leads to the positive effect of stabilisation (for example, the stabilisation of human body temperature is based on negative feedback). The concept of feedback essentially changed our understanding of the cause and effect relationship, resolving paradoxes of circular arguments in logic (when they concern causal reasoning), though it must be understood that such paradoxes can be resolved only by dynamic, not static causal reasoning or models. An example of such paradox is the argument of Bruno Latour [14] against objectivity, saying that since the concept of nature is the outcome of our construction of knowledge, it cannot be at the same time its cause – a clear example of a deep misunderstanding of the essentially dynamic, evolutionary character of the causal positive feedback loop in this case.

But the positive feedback loop between technology and science works slowly: technological stimulations are analyzed by science with much delay, and technology also does not reply instantly to new scientific theories.

The second positive feedback loop is between technology and the systems of its socio-economic applications. The distinction between technology proper and its socio-economic applications should be obvious for at least two reasons. The first is that technologists often work on a technological problem for quite a long time (e.g., almost fifty years in the case of digital television) before their results are broadly socially applied. The second is simple: *technologists do not make much money, technology brokers* (entrepreneurs, managers, bankers, etc.) *do*, just as art brokers make more money than artists. If a technological product or service, such as mobile telephony, produces much revenue, then more money is available for its further technological development; this leads to the truly avalanche-like processes of the social adoption of technological hits.

These processes have strange dynamic properties, socio-economic acceptance of novelties is slow, and there is usually a long delay time between the recognition of a purely technological possibility and the start of an avalanche of its broad socio-economic applications. This delay has many causes which we already discussed, e.g., after initial social distrust, some time must pass before that distrust turns into a *blind social fascination* once a technological hit becomes fashionable. Once it starts to work, the second positive feedback loop is much stronger and faster than the first one. But it can have very dangerous side-effects.

This blind social fascination is actually the autonomous force, incorrectly attributed by social philosophy to technology proper, it is precisely the source of the Heideggerian danger that "man will exalt himself and posture as the lord of the earth".

There are many examples of such blind social fascination. Let us look only at an example of such danger well understood by software specialists, probably missed as yet by philosophers of technology. Consider mobile telephony; the current trend is to integrate many functions in the mobile computer contained actually in any *mobile* (mobile telephony device). One of such functions is global positioning system (GPS); when it is integrated in a mobile, the position on Earth of the user of the mobile can be determined with great accuracy. This has obviously many advantages, and users would pay for this function, mobile manufacturers compete to improve this function, etc., a social fascination. But what if an ambitious minister of interior uses this social fascination to "posture as the lord of the earth" and to implement totalitarian control of people?

[6] Black actually patented this concept in 1928, published a paper on it in 1934.

For such reasons, it is clear that inputs from philosophy of technology might enrich software development and evaluation; the issue, how dangerous is too great accuracy of pinpointing every user of a mobile, cannot be considered as a purely technical or purely economic one, it might require special ethical discussion and legal safeguards. But philosophers of technology, in order to be useful in such a case, should know at which stage of technological development they must participate.

Therefore, let us examine possible stages in some more detail. We list these stages below:

1. Motivation – artistic urge or social demand.
2. Technology proper – actual construction or design of a prototype tool.
3. Testing and evaluation.
4. Transfer from academia to industry.
5. Design of mass production process.
6. Pre-marketing: promotion of demand.
7. Mass production and marketing.
8. Re-engineering based on consumer (user) remarks.
9. Design of new versions due to technological advancement and integration.

The list should not suggest that it is an almost linear[7] process, with few recursions. It is used only to shorten description, while in reality obviously there are many recursions, at each stage different actors are usually involved, multivariate choices have to be made, etc. (see, e.g., [15]).

These nine points only outline the typical stages; some might be omitted, some performed parallel. Stage 1 might be just an idea of a novelty, which I call artistic urge; or perception of future social demand; or realization of actual economic demand, not quite satisfied by market forces. Stage 2 is most important for technology creation and, whether in hardware or software, is not reducible to simple application of hard science; it is artistic, its main motivation is the joy of creation (if an engineer wants to make money, (s)he becomes a manager); this stage is most misunderstood by social constructivists who do not even notice its artistic motivation. Stage 3 is actually equally important: since stages 1 and 2 might be artistic, intuitive, their products must be thoroughly tested and evaluated. Moreover, testing occurs recursively also in other stages, e.g., after preparing a prototype for mass production. In hardware, the tests are very often of destructive character, such as crash tests of cars, just to determine the limits of safe use of new tools. In software, we also often abandon or re-engineer old versions after their tests and evaluation.

Thus, technology is *falsificationist* (in the sense introduced by Karl Popper, see [16]) in its everyday practice. If is a kind of irony, because postmodern sociology of science ridicules falsificationism, saying (with some reasons) that scientists do not try to disprove, only promote their theories; and Karl Popper defended his metaphysical position with regard to the evolution of science, while regarding technology as a mere application of science[8]. Meanwhile, tools are not theories and it is technology that is actually falsificationist, because it is not a mere application of science. This is consistent with (although not noted by) Rachel Laudan [18], who tried to find scientific revolutions of the type of Thomas Kuhn [19] in technology and reported that technological revolutions have quite different character, since technology is more pragmatic, less paradigmatic.

Stage 4 is especially difficult and stages 2 and 3 might be repeated in stage 4; we comment on the reasons in the next section. Naturally, it happens that new technology products (not so often essentially new) are developed directly in industrial laboratories; only then one can speak about factory-like production of knowledge, a favourite theme of postmodern sociology of science. But even then the concept of *techoscience* [14] is a misnomer, because science is paradigmatic, technology falsificationist, they differ essentially in their values and episteme, and in industrial production of knowledge there is a tension between them about the intensity and character of testing and evaluation. Stages 5 and 7 are typical for industrial civilization, and stay important also during informational revolution, though naturally change their character due to automation and robotization, and even more in the case of dematerialized software products. Most of classical philosophy of technology maintained that mass production – particularly Fordism – is the defining characteristics of modern technology, often without noting how information technology made classical Fordism obsolete.

Stage 6 – promotion of demand – must start parallel with preparation of mass production, because it might take more time. Stages 8 and 9 – re-engineering and design of new versions – often proceed parallel.

Now, when should a philosopher of technology participate in this process, to make the outcomes more socially controllable? This corresponds to the question of Langdon Winner [15]: "Where should a philosopher go to learn about technology?"[9] I can accept the argument that it is almost impossible for (her)him to participate in stages 1 and 2; but stage 3 starts relatively early, and the philosopher of technology might be most useful then, also might obtain important insights about the processes of technology creation. Moreover, since (s)he is worried about social

[7] "Linear" in the social science sense of being non-recursive; technologists would rather use "linear" in the sense of linearity of the mathematical model.

[8] Ironically, Karl Popper actually argued that technology does not use falsification [17]. However, his arguments were that engineers do not abandon their designs after falsifying them (obviously not true) and that technology does not use crucial experiments (also not true, much thought is devoted in technology how to devise critical experiments). Thus, his arguments show simply a lack of understanding of technology.

[9] My actual answer is: *philosopher should learn technology as a part of (her) his curriculum at university*, since more than fifty years ago I learned philosophy as a part of mine technological curriculum. But above I propose a less demanding answer.

consequences of technology, (s)he should take also an active part in stage 6.

4. The Difference between Academic and Industrial Knowledge Creation

In order to illustrate the difference of academic and industrial knowledge creation, I review here some results concerning so-called micro-theories of knowledge creation. The demands of knowledge based economy resulted recently in the emergence of many such micro-theories – concerning knowledge creation for the needs of today and tomorrow, as opposed to classical concentration of philosophy on macro-theories of knowledge creation on a long term historical scale. Historically, we could count the concept of *brainstorming* [20] as an early example of such micro-theories. Since 1990 we observe many such new micro-theories originating in systems science, management science and information science, beginning with the *Shinayakana systems approach* [21], the *knowledge creating company* and the *SECI* (socialization-externalization-combination-internalization) *spiral* [22], the *rational evolutionary theory of intuition* [23], the I^5 *(pentagram) system* [24], the *OPEC* (objectives-process-expansion-closure) *spiral* [25] and several others. All such recent micro-theories of knowledge creation processes take into account the interplay of tacit, intuitive, emotive, and preverbal aspects with explicit or rational aspects of knowledge creation, as well as the interplay between an individual and a group.

Additional results concerning micro-theories of knowledge creation were obtained also in the 21st Century COE Program Technology Creation Based on Knowledge Science at the Japan Advanced Institute of Science and Technology (JAIST). For example, the brainstorming process was represented as the *DCCV* (divergence-convergence-crystallization-verification) *spiral* [26] due to the research in this program. The concept of *creative space* [4] tries to provide a synthesis of such diverse micro-theories.

We shall not discuss here in detail the rational evolutionary *theory of powerful but fallible intuition* (see [4], [23], [27]). The introduction of a three-by three matrix *rational-intuitive-emotive* and *individual-group-humanity* knowledge used in [4] instead of two-by-two *explicit-tacit* and *individual-group* used as the basis of the *SECI spiral* in [22] makes it possible to generalize the *SECI spiral* into a network-like model of creative processes, called *creative space* (see Fig. 2).

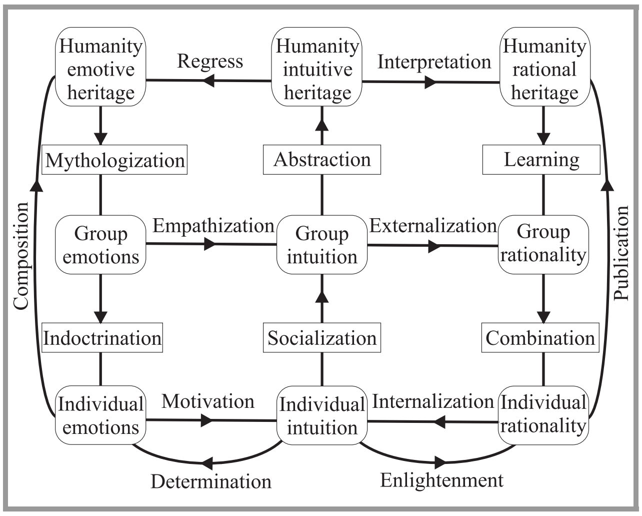


Fig. 2. Basic dimensions of creative space.

The model of *creative space* consists of *nodes* – such as *Individual rationality* or individual rational knowledge – and *transitions*[10] between the nodes — such as *Internalization* from *Individual rationality* to *Individual intuition*. Note that the *SECI spiral* of [22] is essentially preserved in the lower right-hand corner of Fig. 2; but creative space involves also many other transitions. For example, the upper left-hand corner of Fig. 2 represents [28] the theory of revolutionary scientific change in the form of the *ARME* (abstraction-regress-mythologization-emphatization) *spiral*, see [4] for a more detailed discussion.

Other dimensions can be added to the model of creative space and many other knowledge creation processes can be represented in the model. Knowledge management is naturally more interested in the processes of *normal* knowledge creation (as opposed to *revolutionary*; this distinction is due to Kuhn [19]). In [5], two types of normal knowledge creation processes are distinguished:

- *Organizational or industrial* processes in market or purpose-oriented knowledge creation, such as the *SECI spiral* of Nonaka and Takeuchi. Such processes are motivated mostly by the interests of a group and two other spirals of this type can be also represented in creative space; these are the *brainstorming DCCV spiral* [26] and the occidental counterpart of *SECI spiral*, the *OPEC spiral* of [25].

- *Academic* processes of normal knowledge creation, in universities and research institutes. Such processes are motivated mostly by the interests of an individual researcher. Three typical spirals of this type are distinguished as parts of *creative space* in [4]:
  - the *hermeneutic* (enlightenment-analysis-hermeneutic immersion-reflection) *EAIR spiral* of reading and interpreting scientific literature;
  - the *debating EDIS* (enlightenment-debate-immersion-selection) *spiral* of scientific discussions;
  - the *experimental EEIS* (enlightenment-experiment-interpretation-selection) *spiral* of performing experiments and interpreting their results.

[10] Originally called *conversions* by Nonaka and Takeuchi [22], but knowledge is not lost when used, hence it cannot be converted; thus we prefer the more neutral term *transitions*.

We should note that all these three spirals begin with the transition *enlightenment* from *individual intuition* to *individual rationality* (called also variously *abduction*, *aha*, *eureka*, *illumination* – simply having an idea – and indicated in the bottom right-hand part of Fig. 2). Because of that, we can switch between these three spirals or perform them parallel. This is indicated in Fig. 3, where these three spirals are presented together as a *triple helix* of normal academic knowledge creation.

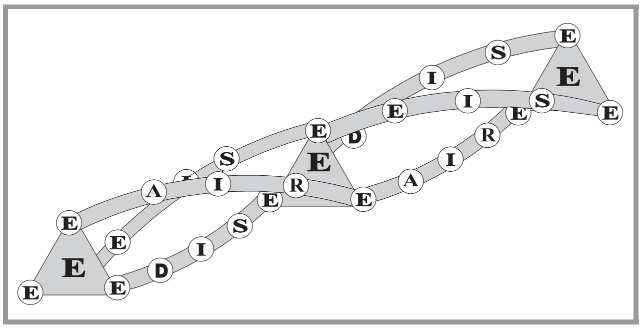


Fig. 3. The triple helix of normal academic knowledge creation.

Thus, academic knowledge creation processes are quite different than organizational knowledge creation; understanding their differences might help in overcoming the difficulty of cooperation between academia and industry. Alternatively, we could try to combine them, see below.

These three spirals contained in the triple helix do not exhaustively describe all what occurs in academic knowledge creation, but they describe most essential elements of academic research: gathering and interpreting information and knowledge, debating and experimenting. In fact, recent research including a questionnaire on creativity conditions in JAIST supported, both directly and indirectly, the conclusion that these elements are very important for academic knowledge creation, see [4], [29]. However, these spirals are *individually oriented*, even if a university and a laboratory should support them, e.g., the motivation for and the actual research on preparing a doctoral thesis is mostly individual. Moreover, the triple helix only describes what researchers actually do, it is thus a *descriptive* model. Obviously, the model helps in a better understanding of some intuitive transitions in these spirals and makes possible testing, which parts of these spirals are well supported in academic practice and which require more support; but it does not give clear conclusions *how to organize research*.

However, the three spirals of organizational knowledge creation mentioned before are important for practical knowledge creation, for innovations, particularly in industry and other purpose-oriented organizations. Unfortunately, they cannot be easily combined into a multiple helix like the triple helix, because they do not share the same elements. The main challenge is not only to combine these spirals between themselves, but also with the spirals of academic knowledge creation. This general challenge is difficult, but such a combination would be important for several reasons:

– combining these spirals might strengthen academic knowledge creation, because it would increase the role of the group supporting the individual research;

– combining these spirals might strengthen also industrial innovation and knowledge creation, because it always contains also some individual elements that should be explicitly accounted for;

– combining these spirals might help in the cooperation of industry with academic institutions in producing innovations, because it could bridge the gap between the different ways of conducting research in academia and in industry.

The idea of *Nanatsudaki model of knowledge creation processes* [5] tries to derive pragmatic conclusions from such analysis and synthesis, by combining seven spirals (objective setting OPEC, hermeneutic EAIR, socializing SECI, brainstorming DCCV, debating EDIS, roadmapping I-System [30], and experimenting EEIS) in an order useful for organizing large research projects, particularly for cooperation between academia and industry, see Fig. 4.

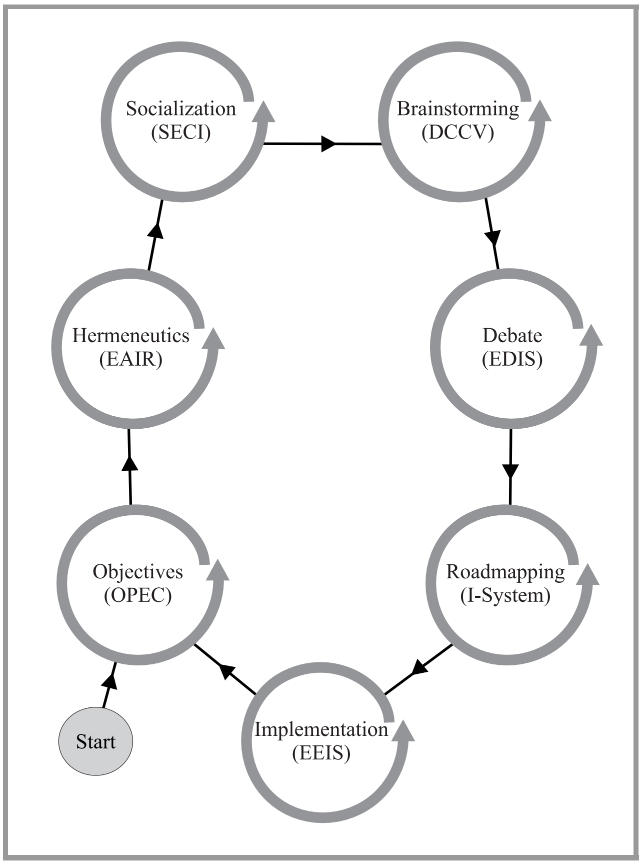


Fig. 4. Diagram of JAIST Nanatsudaki model (septagram of creative spirals).

The conclusion from this short discussion of the differences between academic and industrial knowledge creation is clear: they are essentially different, both by the difference between individual and group orientation and by the different character of typical processes used. We might try to overcome these differences, both in praxis and by suggesting models such as *Nanatsudaki septagram*, but overcoming these differences causes additional delays in technology development.

5. Changing Episteme

We understand here the concept of *episteme* in the sense of [31] – as *the way of creating and justifying knowledge characteristic for a given historical era*. However, if we live in time of a change of historical eras, then we should also observe a process of a change of the episteme. How does such process proceed?

During last stages of a former era, a change of the episteme is prepared by an accumulation of new concepts inconsistent with the old episteme and a following destruction of this old episteme. The new concepts inconsistent with the *modern episteme* that accumulated during the 20th century were, between others:

- relativity and relativism;
- indetermination and pluralism;
- feedback and dynamic systemic development;
- deterministic and probabilistic chaos, order emerging out of chaos;
- butterfly effect and change;
- complexity and emergence principle;
- computational complexity as a limit on cognitive power;
- logical pluralism;
- new theories of knowledge creation, etc.

This has led to a destruction of the modern episteme; in the second half of the 20th century, such a destruction resulted in a divergent development of the episteme of three cultural spheres of:

- basic, hard and natural sciences;
- social sciences and humanities;
- technology.

Thus, we should speak not about *two cultures* [32], but about *three distinct episteme* [11].

These cultural spheres adhere to different values, use different concepts and languages, follow different paradigms or underlying them hermeneutical horizons; *such differences increased gradually with the development of poststructuralism and postmodernism, while hard sciences and technology went quite different epistemic ways*.

Obviously, *technology cooperates strongly with hard and natural sciences*, as shown in Fig. 1; but there is an essential epistemic difference between these two spheres: hard and natural sciences are paradigmatic, while technology is not paradigmatic, rather pragmatic. However, both hard sciences and technology know for a long time (e.g., since [33]) that knowledge is constructed by humans, only they interpret this diversely.

Even if a hard scientist knows that all knowledge is constructed and there are no absolute truth and objectivity, he believes that scientific theories are *laws of nature discovered by humans* rather than *models of knowledge created by humans*. He values truth and objectivity as ultimate ideals; metaphorically, hard scientist resembles a priest.

A technologist is much more *relativist and pragmatic* in his episteme, he readily agrees that *scientific theories are models of knowledge*; but requires that these theories should be *as objective as possible*, tested in practice, he demands that they should be *falsifiable* (as postulated by Karl Popper [16]). Metaphorically, a *technologist resembles an artist* (see also [10]–[12]), also values tradition like an artist does, much more than a scientist.

Without discussing in further detail the observed differences between the episteme of these three cultural spheres and many actual examples of these epistemic differences – *science wars* in the last decade of the 20th century were a clear indication of them (see, e.g., [34]), we turn to other conclusions.

If we are living in times of an *informational revolution* and this revolution leads to a new civilization era, in which knowledge plays an even more important role than just information – thus we might call the new epoch the *knowledge civilization era* – then we might expect a formation of a new, integrated episteme characteristic for the new era. This usually occurs, as shown by Foucault [31], after the beginning of the new era. Naturally, the dates of beginnings of historical eras are conventional, reflect a given hermeneutical horizon, but are best defined by historians. Thus we follow the example of Fernard Braudel [35] who defined the long duration preindustrial era of the beginnings of capitalism, of print and geographic discoveries, as starting in 1440 with Gutenberg, and ending in 1760 with Watt. Following his example, we *select 1980* – the time when information technology was made broadly socially available by the introduction of personal computers and computer networks – *as the beginning date of the new era of knowledge civilization*, even though computers were used earlier[11].

[11] All these turning points were not new inventions, but improvements of older inventions that enabled, however, their broad social use. Print was known in China before Gutenberg, but it was inefficient, could not result in a mass production of books. Steam engine was known before Watt, but it tended to explode, could not be used broadly. Computers were known before Apple Co. produced first personal computer, but former computers were giant machines used only by specialists.

Thus, instead of three waves of [36] we speak about recent *three civilization eras*:

- *preindustrial civilization* (*formation of capitalism*) 1440–1760;
- *industrial civilization* 1760–1980;
- *knowledge civilization* 1980–2100+(?)

The date 2100+ means "at least until 2100" and is not only a simple prediction based on shortening periods of these eras (320–220–120?), it can be substantiated also differently, see [6].

The new episteme characteristic for the era of knowledge civilization must be an integration of the diverged epistemic positions of hard sciences, soft sciences and technology. An attempt of such integration was made, e.g., in [5]. However, here we shall quote only some elements that might help in the integration of the new episteme, namely, the *multimedia principle* and the *emergence principle*. These two principles were first formulated in [4], [5].

Multimedia principle: words are just an approximate code to describe a much more complex reality, visual and preverbal information in general is much more powerful and relates to intuitive knowledge and reasoning; the future records of the intellectual heritage of humanity will have a multimedia character, thus stimulating creativity.

Emergence principle: new properties of a system emerge with increased levels of complexity, and these properties are qualitatively different than and irreducible to the properties of its parts.

Both these principles might seem to be just common sense, intuitive perceptions; the point is that they are justified rationally and scientifically. Moreover, they go beyond and are in a sense opposed to fashionable trends in poststructuralism and the postmodern philosophy or sociology of science.

The multimedia principle is based on the technological and information science knowledge: as shown in the rational evolutionary theory of intuition [23], *a figure is worth at least ten thousand words*. The poststructuralist philosophy stresses the roles of *metaphors and icons*, but reduces them to *signs*, which is contrary to the emergence principle. Thus, the world is not constructed by us in a social discourse, as the poststructuralist and postmodern philosophy wants us to believe: *we observe the world by all our senses, including vision, and strive to find adequate words* when trying to describe our preverbal impressions and thinking *to communicate them in language*. Language is a shortcut in civilization evolution of humans, our original thinking is preverbal, often unconscious.

Multimedia principle originates in technology and has diverse implications for technology creation. *Information technology creation should concentrate on multimedia aspects of supporting communication and creativity. Technology creation starts essentially with preverbal thinking.*

The emergence principle is also partly motivated by technological experience. It stresses that new properties of a system emerge with increased levels of complexity, and *these properties are qualitatively different than and irreducible* to the properties of its parts. This might appear to be just a conclusion from the classical concepts of systems science, synergy and holism; or just a metaphysical religious belief. The point is that both such simplifying conclusions are mistaken. *Synergy and holism say that a whole is greater than the sum of its parts, but do not stress irreducibility. Thus, according to classical systemic reasoning, a whole is greater, but still explicable by and reducible to its parts.*

The best recent example of the phenomenon of emergence is the concept of software that spontaneously emerged in the civilization evolution during last fifty years. *Software cannot function without hardware, but is irreducible to and cannot be explained by hardware*. This has also some importance for the metaphysics of the absolute, because it is also a negation of the arguments of creationists who say that irreducible complexity could not emerge spontaneously in evolution.

Both multimedia principle and emergence principle might be interpreted as having some metaphysical character, in the same sense as Karl Popper admitted [16] that his falsification principle has metaphysical character. The emergence principle, however, is not a metaphysical religious belief, because it can be justified rationally and scientifically (see [4], [5]) – even if it might have serious metaphysical consequences.

Based on the concepts presented above, we might turn back to the issue of basic explanations of development of science and technology. People, motivated by curiosity and aided by intuition and emotions, formulate hypotheses about properties of nature and of human relations; they also construct tools that help them to deal with nature or with other people; together, we call all this knowledge (see also [37]). People test and evaluate the knowledge constructed by them by applying it to reality: perform destructive tests of tools, devise critical empirical tests of theories concerning nature, apply and evaluate theories concerning social and economic relations. Such a process can be represented as a general spiral of evolutionary knowledge creation [5], [34], see Fig. 5.

We observe reality and its changes, compare our observations with human intellectual heritage (*Observation*). Then our intuitive and emotive knowledge helps us to generate new knowledge (*Enlightenment*); we apply new knowledge to existing reality (*Application*), obtain some changes of reality (*Modification*). We observe them again and modified reality becomes existing reality through *Recourse*; only the positively tested knowledge, resilient to falsification attempts, remains an important part of human heritage (*Evaluation*); this can be interpreted as an objectifying, stabilizing feedback.

Thus, *nature is not only the effect of construction of knowledge by people, nor it is only the cause of knowledge: it is both cause and effect in a positive feedback loop, where more knowledge results in more modifications of nature and*

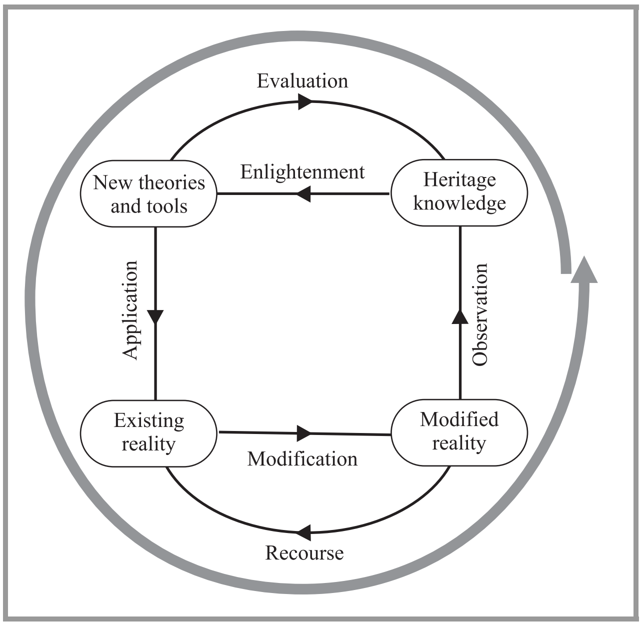


Fig. 5. The general OEAM spiral of evolutionary knowledge creation.

more modifications result in more knowledge. The overall result is an avalanche-like growth of knowledge, although it can have slower normal and faster revolutionary periods. This avalanche-like growth, if unchecked by stabilizing feedbacks, beside tremendous opportunities creates also diverse dangers, usually not immediately perceived but lurking in the future.

Moreover, we should select knowledge that is as objective as possible (even if absolute objectivity is impossible; which we know since Heisenberg [38]) because avalanche-like growth creates diverse threats: we must leave to our children best possible knowledge in order to prepare them for dealing with unknown future.

6. What New Topics Can We List for Philosophy of Technology

Science, particularly soft science, is paradigmatic; this concerns also philosophy of technology, no matter whether we would classify philosophy as science or as a separate part of human knowledge. *Philosophy of technology* follows certain exemplars: it is general, sees technology as a socio-economic system of producing and utilizing products of technology; sometimes accuses this system of being uncontrollable, or developing without taking into account humanistic values; or even being unethical; is historically oriented, concentrates on *ex post* evaluation of results of technological development – and *avoids learning details of technological development*. This concerns even the most technology-friendly, new approaches to philosophy of technology, such as Don Ihde [39] or Peter-Paul Verbeek [40].

We have seen that this paradigm is self-reinforcing: if you don't want to learn details of a process and the process has internal, large delays, then you will consider this process uncontrollable; if you judge that some process is uncontrollable, then you do not think it is worth learning its' details. We have seen that this paradigm is just opposite to the praxis of software development and evaluation which – despite inevitable delays – concentrates practically on the control of an important technological process.

Therefore, a way of changing the paradigmatic attitude of philosophy of technology is to invite some such philosophers to participate in software development and evaluation; if they want to have an impact on contemporary technology development, to a large extent determined by software, it is an invitation they could not refuse without risking becoming outdated and sterile. And they can enrich software evaluation, contribute to its' ethical and sociological dimensions.

However, when participating in software evaluation, philosophers of technology will inevitably learn more about software, about its evaluation methods, about the falsificationist approach of technology to knowledge creation – thus they will be forced to revise their paradigm. Thus, inviting philosophers to participate in software evaluation is not only in the interest of the latter; it is also in the interest of better understanding the character of informational revolution and knowledge civilization era.

If the philosophers of technology learn more about informational revolution, they might be also helpful in resolving new conflicts and counteracting new dangers related to the knowledge civilization era. We shall discuss here shortly only two such issues.

First is *the conflict between the individual property of knowledge, supported by public property of the human intellectual heritage, and the corporate property of knowledge* that attempts to subjugate the individual knowledge of employees and to privatise the human intellectual heritage. This conflict is inevitable if knowledge becomes the main productive resource. Old arguments supporting privatization (e.g., so called tragedy of commons: common use of a pasture leads to a degradation of the common resource) are used to justify the actions of the corporate side, but knowledge has different properties than other productive resources, it is not degradable, its use results in an increase, not a decrease of the jointly held knowledge. Thus, it is better for society if knowledge remains public resource, while many corporate companies realized long ago that it is in their interest to privatize this public property. This conflict is already perceived, see [41], and academy strikes back by promoting the idea of open access and demanding that all results of research financed by public resources should be freely accessible in net portals. However, this is only the beginning of a conflict that might become the defining one for the knowledge civilization era.

The second is an *actual threat of computer and robot domination of people*. Such a threat was a natural subject of science fiction; on the other hand, some excellent books, e.g., [42], have shown that human mind cannot be duplicated by a classical computer intelligence. However, while the rational evolutionary theory of intuition [23] supports such a conclusion, it also shows that the difference, although tremendous, is quantitative (e.g., a figure is worth at least ten thousand words). Thus, if the Moore law [43] holds for another two or three decades, new aspects of the intelligence of computers might emerge and it might start to be comparable even to the intuitive and emotive intelligence of humans. Moreover, already now robots are being used as weapons; what about their use by terrorists or organized crime? Thus, we should not regard this issue as a distant, science fiction question, but already now start debates how to counteract these dangers.

7. Conclusions

The main points of this paper are:

- Technology development process is complex, has many stages and includes many delays, in historical evidence amounting sometimes to fifty years.
- When seen holistically from outside, such a process is apt to appear uncontrollable (called also autonomous or deterministic by philosophy of technology) – but this is only a matter of perspective.
- When seen from inside and in technological detail, such a process is very much controllable, and several technological disciplines concentrate on the ways of controlling such processes.
- Software development and evaluation is one of such processes, very important for future technology development.
- It is worthwhile to invite philosophers of technology to participate in software evaluation; this might even contribute to the change of the paradigm of philosophy of technology.
- Philosophy of technology should be also aware of new conflicts and dangers related to the knowledge civilization era.

A general conclusion is also that we need paradigm changes and essentially new approaches to many issues, such as software development and evaluation versus philosophy of technology, in the time when informational revolution results in a transition towards knowledge civilization.

References

[1] V. Dusek, *Philosophy of Technology*. Oxford: Blackwell Publ., 2006.

[2] A. Bard and J. Söderqvist, *Netokracja. Nowa elita władzy i życie po kapitalizmie*. Warsaw: Wydawnictwa Akademickie i Profesjonalne, 2006 (Polish transl.).

[3] Z. Król, “The emergence of new concepts in science”, in *Creative Environments: Supporting Creativity for the Knowledge Civilization Age*, A P. Wierzbicki and Y. Nakamori, Eds. Berlin-Heidelberg: Springer-Verlag, 2007.

[4] A. P. Wierzbicki and Y. Nakamori, *Creative Space: Models of Creative Processes for the Knowledge Civilization Age*. Berlin-Heidelberg: Springer-Verlag, 2006.

[5] A. P. Wierzbicki and Y. Nakamori, *Creative Environments: Supporting Creativity for the Knowledge Civilization Age*. Berlin-Heidelberg: Springer-Verlag, 2007.

[6] A. Kameoka and A. P. Wierzbicki, “A vision of new era of knowledge civilization”, in *Proc. Ith World Congr. IFSR*, Kobe, Japan, 2005.

[7] A. P. Wierzbicki, “Prinzip maksimuma dla processov s nietrivialnom zapazdiwaniyem upravleniya” (Maximum principle for processes with non-trivial delay of control), *Avtomatika i Telemekhanika*, no. 10, 1970 (in Russian).

[8] J. Ellul, *The Technological Society*. New York: Alfred Knopf, 1964.

[9] C. Mitcham, *Thinking through Technology: The Path between Engineering and Philosophy*. Chicago: The University of Chicago Press, 1994.

[10] M. Heidegger, “Die Technik und die Kehre”, in M. Heidegger, *Vorträge und Aufsätze*. Pfullingen: Günther Neske Verlag, 1954.

[11] A. P. Wierzbicki, “Technology and change: the role of technology in knowledge civilization”, in *Proc. Ith World Congr. IFSR*, Kobe, Japan, 2005.

[12] A. P. Wierzbicki, “Technology and change: the role of information technology in knowledge civilization”, *J. Telecommun. Inform. Technol.*, no. 4, pp. 3–14, 2006.

[13] H. S. Black, “Stabilized feedback amplifiers”, *Bell Syst. Tech. J. Electr. Eng.*, vol. 53, pp. 1311–1312, 1934.

[14] B. Latour, *Science in Action*. Milton Keynes: Open University Press, 1987.

[15] L. Winner, “Social constructivism: opening the black box and finding it empty”, *Sci. Cult.*, vol. 16, pp. 427–452, 1993.

[16] K. R. Popper, *Objective Knowledge*. Oxford: Oxford University Press, 1972.

[17] K. R. Popper, “Three views concerning human knowledge”, in *Contemporary British Philosophy*, *Third Series*, H. D. Lewis, Ed. London: Allen and Unwin, 1956.

[18] *The Nature of Technological Knowledge. Are Models of Scientific Change Relevant?* R. Laudan, Ed. Dordrecht: Reidel, 1984.

[19] T. S. Kuhn, *The Structure of Scientific Revolutions*. Chicago: Chicago University Press, 1962 (2nd ed., 1970).

[20] A. F. Osborn, *Applied Imagination*. New York: Scribner, 1957.

[21] Y. Nakamori and Y. Sawaragi, “Shinayakana systems approach in environmental management”, in *Proceedings of 11th World Congress of International Federation of Automatic Control*. Tallin: Pergamon Press, 1990, vol. 5, pp. 511–516.

[22] I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company. How Japanese Companies Create the Dynamics of Innovation*. New York: Oxford University Press, 1995.

[23] A. P. Wierzbicki, “On the role of intuition in decision making and some ways of multicriteria aid of intuition”, *Multiple Crit. Decis. Mak.*, vol. 6, pp. 65–78, 1997.

[24] Y. Nakamori, “Knowledge management system toward sustainable society”, in *Proc. First Int. Symp. Knowl. Syst. Sci.*, Tatsunokuchi, Japan, 2000, pp. 57–64.

[25] S. Gasson, “The management of distributed organizational knowledge”, in *Proceedings of the 37 Hawaii International Conference on Systems Sciences (HICSS 37)*, R. J. Sprague, Ed. Los Alamitos: IEEE Computer Society Press, 2004.

[26] S. Kunifuji, “Creativity support systems in JAIST”, in *Proc. JAIST Forum 2004: Technol. Creat. Bas. Knowl. Sci.*, Ishikawa, Japan, 2004, pp. 56–58.

[27] A. P. Wierzbicki, "Knowledge creation theories and rational theory of intuition", *Int. J. Knowl. Syst. Sci.*, vol. 1, pp. 17–25, 2004.

[28] A. Motycka, *Nauka a nieświadomość* (Science and Unconscious). Wrocław: Leopoldinum, 1998 (in Polish).

[29] J. Tian, A. P. Wierzbicki, H. Ren, and Y. Nakamori, "A study on knowledge creation support in a Japanese research institute", *Int. J. Knowl. Syst. Sci.*, vol. 3, no. 1, pp. 7–17, 2006.

[30] T. Ma, S. Liu, and Y. Nakamori, "Roadmapping for supporting scientific research", in *Proc. MCDM 2004, 17th Int. Conf. Multiple Crit. Decis. Mak.*, Whistler, Canada, 2004.

[31] M. Foucault, *The Order of Things: an Archaeology of Human Sciences*. New York: Routledge, 1972.

[32] C. P. Snow, T*he Two Cultures*. Cambridge: Cambridge University Press, 1960.

[33] W. V. Quine, "Two dogmas of empiricism", in *Philosophy of Mathematics*, P. Benacerraf and H. Putnam, Eds. Englewood Cliffs: Prentice-Hall, 1964.

[34] A. P. Wierzbicki, "Group decisions and negotiation in the knowledge civilization era", in *Proc. 9th Group Decis. Negotiat. Conf.*, Coimbra, Portugal, 2008.

[35] F. Braudel, *Civilisation matérielle, économie et capitalisme, XV–XVIII siècle*. Paris: Armand Colin, 1979.

[36] A. Toffler and H. Toffler, *The Third Wave*. New York: William Morrow, 1980.

[37] H. S. Jensen, L. M. Richter, and M. T. Vendelø, *The Evolution of Scientific Knowledge*. Cheltenham: Edward Elgar, 2003.

[38] W. Heisenberg, "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik", *Zeitschrift für Physik*, vol. 43, pp. 172–198, 1927.

[39] D. Ihde, *Bodies in Technology*. Minneapolis – London: University of Minnesota Press, 2002.

[40] P. P. Verbeek, *What Things Do*. University Park: The Pennsylvania University Press, 2005.

[41] L. Lessig, *Free Culture: The Nature and Future of Creativity*. London: Penguin Books, 2004.

[42] H. Dreyfus and S. Dreyfus, *Mind over Machine: The Role of Human Intuition and Expertise in the Era of Computers*. New York: Free Press, 1986.

[43] G. A. Moore, "Cramming more components onto integrated circuits", *Electronics*, vol. 38, no. 8, 1965.

Andrzej Piotr Wierzbicki got his M.Sc. in 1960 in telecommunications and control engineering, Ph.D. in 1964 in nonlinear dynamics in control, and D.Sc. in 1968 in optimization and decision science. He worked as the Dean of the Faculty of Electronics, Warsaw University of Technology (WUT), Poland (1975–1978); the Chairman of Systems and Decision Sciences Program of International Institute for Applied Systems Analysis in Laxenburg n. Vienna, Austria (1979–1984). He was elected a member of the State Committee for Scientific Research of Republic of Poland and the Chairman of its Commission of Applied Research (1991–1994). He was the Director General of the National Institute of Telecommunications in Poland (1996–2004). He worked as a research professor at Japan Advanced Institute of Science and Technology (JAIST), Nomi, Ishikawa, Japan (2004–2007). Beside teaching and lecturing for over 45 years and promoting over 100 master's theses and 20 doctoral dissertations at WUT, he also lectured at doctoral studies at many Polish and international universities. Professor Wierzbicki is an author of over 200 publications, including 14 books, over 80 articles in scientific journals, over 100 papers at conferences; the author of 3 patents granted and industrially applied. His current interests include vector optimization, multiple criteria and game theory approaches, negotiation and decision support, issues of information society and knowledge civilization, rational evolutionary theory of intuition, theories of knowledge creation and management.

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

Invited paper

Reference Point Method with Importance Weighted Partial Achievements

Włodzimierz Ogryczak

Abstract—The reference point method (RPM) is based on the so-called augmented max-min aggregation where the worst individual achievement maximization process is additionally regularized with the average achievement. In order to avoid inconsistencies caused by the regularization, we replace it with the ordered weighted average (OWA) which combines all the individual achievements allocating the largest weight to the worst achievement, the second largest weight to the second worst achievement, and so on. Further following the concept of the weighted OWA (WOWA) we incorporate the importance weighting of several achievements into the RPM. Such a WOWA RPM approach uses importance weights to affect achievement importance by rescaling accordingly its measure within the distribution of achievements rather than by straightforward rescaling of achievement values. The recent progress in optimization methods for ordered averages allows us to implement the WOWA RPM quite effectively as extension of the original constraints and criteria with simple linear inequalities.

Keywords— aggregation methods, multicriteria decision making, reference point method, WOWA.

1. Introduction

Consider a decision problem defined as an optimization problem with m criteria (objective functions). In this paper, without loss of generality, it is assumed that all the criteria are maximized (that is, for each outcome "more is better"). Hence, we consider the following multiple criteria optimization problem:

$$\max \{ \ (f_1(\mathbf{x}),\ldots,f_m(\mathbf{x})) \ : \ \mathbf{x} \in Q \ \}, \qquad (1)$$

where $\mathbf{x}$ denotes a vector of decision variables to be selected within the feasible set $Q \subset R^n$, and $\mathbf{f}(x) = (f_1(\mathbf{x}), f_2(\mathbf{x}),\ldots,f_m(\mathbf{x}))$ is a vector function that maps the feasible set Q into the criterion space R^m. Note that neither any specific form of the feasible set Q is assumed nor any special form of criteria $f_i(\mathbf{x})$ is required. We refer to the elements of the criterion space as outcome vectors. An outcome vector $\mathbf{y}$ is attainable if it expresses outcomes of a feasible solution, i.e., $\mathbf{y} = \mathbf{f}(\mathbf{x})$ for some $\mathbf{x} \in Q$.

Model (1) only specifies that we are interested in maximization of all objective functions f_i for $i \in I = \{1,2,\ldots,m\}$. Thus it allows only to identify (to eliminate) obviously inefficient solutions leading to dominated outcome vectors, while still leaving the entire efficient set to look for a satisfactory compromise solution. In order to make the multiple criteria model operational for the decision support process, one needs assume some solution concept well adjusted to the decision maker (DM) preferences. This can be achieved with the so-called quasi-satisficing approach to multiple criteria decision problems. The best formalization of the quasi-satisficing approach to multiple criteria optimization was proposed and developed mainly by Wierzbicki [1] as the reference point method (RPM). The reference point method was later extended to permit additional information from the DM and, eventually, led to efficient implementations of the so-called aspiration/reservation based decision support (ARBDS) approach with many successful applications [2]–[5].

The RPM is an interactive technique. The basic concept of the interactive scheme is as follows. The DM specifies requirements in terms of reference levels, i.e., by introducing reference (target) values for several individual outcomes. Depending on the specified reference levels, a special scalarizing achievement function is built which may be directly interpreted as expressing utility to be maximized. Maximization of the scalarizing achievement function generates an efficient solution to the multiple criteria problem. The computed efficient solution is presented to the DM as the current solution in a form that allows comparison with the previous ones and modification of the reference levels if necessary.

The scalarizing achievement function can be viewed as two-stage transformation of the original outcomes. First, the strictly monotonic partial achievement functions are built to measure individual performance with respect to given reference levels. Having all the outcomes transformed into a uniform scale of individual achievements they are aggregated at the second stage to form a unique scalarization. The RPM is based on the so-called augmented (or regularized) max-min aggregation. Thus, the worst individual achievement is essentially maximized but the optimization process is additionally regularized with the term representing the average achievement. The max-min aggregation guarantees fair treatment of all individual achievements by implementing an approximation to the Rawlsian principle of justice.

The max-min aggregation is crucial for allowing the RPM to generate all efficient solutions even for nonconvex (and particularly discrete) problems. On the other hand, the regularization is necessary to guarantee that only efficient solution are generated. The regularization by the average achievement is easily implementable but it may disturb the basic max-min model. Actually, the only consequent regularization of the max-min aggregation is the lex-min order or more practical the ordered weighted averages (OWA) aggregation with monotonic weights. The latter combines all the partial achievements allocating the largest weight to the worst achievement, the second largest weight to the second worst achievement, the third largest weight to the third worst achievement, and so on. The recent progress in optimization methods for ordered averages [6] allows one to implement the OWA RPM quite effectively. Further following the concept of weighted OWA [7] the importance weighting of several achievements may be incorporated into the RPM. Such a weighted OWA (WOWA) enhancement of the RPM uses importance weights to affect achievement importance by rescaling accordingly its measure within the distribution of achievements rather than straightforward rescaling of achievement values [8]. The paper analyzes both the theoretical and implementation issues of the WOWA enhanced RPM.

2. Scalarizations of the RPM

While building the scalarizing achievement function the following properties of the preference model are assumed. First of all, for any individual outcome y_i more is preferred to less (maximization). To meet this requirement the function must be strictly increasing with respect to each outcome. Second, a solution with all individual outcomes y_i satisfying the corresponding reference levels is preferred to any solution with at least one individual outcome worse (smaller) than its reference level. That means, the scalarizing achievement function maximization must enforce reaching the reference levels prior to further improving of criteria. Thus, similar to the goal programming approaches, the reference levels are treated as the targets but following the quasi-satisficing approach they are interpreted consistently with basic concepts of efficiency in the sense that the optimization is continued even when the target point has been reached already.

The generic scalarizing achievement function takes the following form [1]:

$$S(\mathbf{y}) = \min_{1\le i\le m} \{s_i(y_i)\} + \frac{\varepsilon}{m}\sum_{i=1}^{m} s_i(y_i)\,, \tag{2}$$

where ε is an arbitrary small positive number and $s_i : R \rightarrow R$, for $i = 1,2,\ldots,m$, are the partial achievement functions measuring actual achievement of the individual outcomes y_i with respect to the corresponding reference levels. Let a_i denote the partial achievement for the ith outcome ($a_i = s_i(y_i)$) and $\mathbf{a} = (a_1,a_2,\ldots,a_m)$ represent the achievement vector. The scalarizing achievement function (2) is, essentially, defined by the worst partial (individual) achievement but additionally regularized with the sum of all partial achievements. The regularization term is introduced only to guarantee the solution efficiency in the case when the maximization of the main term (the worst partial achievement) results in a non-unique optimal solution. Due to combining two terms with arbitrarily small parameter ε, formula (2) is easily implementable and it provides a direct interpretation of the scalarizing achievement function as expressing utility.

Various functions s_i provide a wide modeling environment for measuring partial achievements [5], [9]–[11]. The basic RPM model is based on a single vector of the reference levels, the aspiration vector $\mathbf{r}^a$. For the sake of computational simplicity, the piecewise linear functions s_i are usually employed. In the simplest models, they take a form of two segment piecewise linear functions:

$$s_i(y_i) = \begin{cases} \lambda_i^+(y_i - r_i^a), & \text{for} \quad y_i \ge r_i^a \\ \lambda_i^-(y_i - r_i^a), & \text{for} \quad y_i < r_i^a\,, \end{cases} \tag{3}$$

where $\lambda_i^+ > \lambda_i^-$ are positive scaling factors corresponding to underachievements and overachievements, respectively, for the ith outcome. It is usually assumed that λ_i^+ is much larger than λ_i^-. Figure 1 depicts how differentiated scaling affects the isoline contours of the scalarizing achievement function.

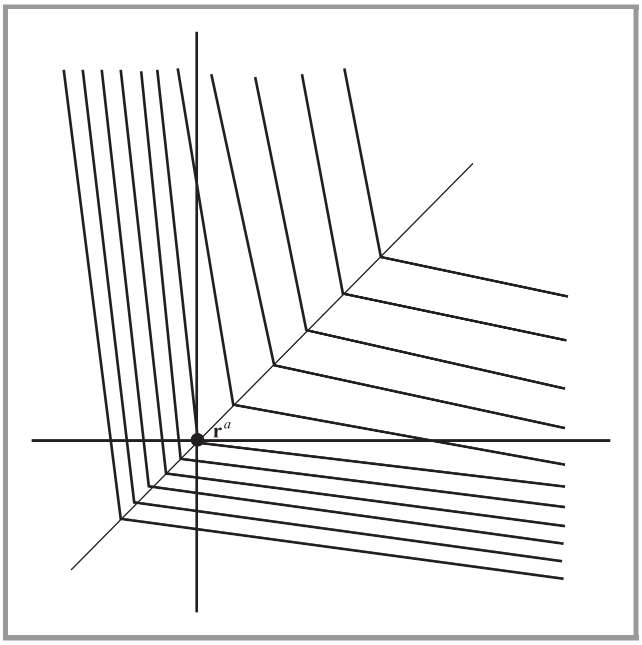


Fig. 1. Isoline contours for the scalarizing achievement function (2) with partial achievements (3).

Real-life applications of the RPM methodology usually deal with more complex partial achievement functions defined with more than one reference point [5] which enriches the preference models and simplifies the interactive analysis. In particular, the models taking advantages of two reference vectors: vector of aspiration levels $\mathbf{r}^a$ and vector of reservation levels $\mathbf{r}^r$ [3] are used, thus allowing the DM to specify requirements by introducing acceptable and required values for several outcomes. The partial achievement function s_i can be interpreted then as a measure of the DM's satisfaction with the current value of outcome the ith criterion. It is a strictly increasing function of outcome y_i with value $a_i = 1$ if $y_i = r_i^a$, and $a_i = 0$ for $y_i = r_i^r$. Thus the partial achievement functions map the outcomes values onto a normalized scale of the DM's satisfaction. Various functions can be built meeting those requirements. We use the piece-wise linear partial achievement function introduced in an implementation of the ARBDS system for the multiple criteria transshipment problems with facility location [12]:

$$s_i(y_i) = \begin{cases} \gamma \dfrac{y_i - r_i^r}{r_i^a - r_i^r}, & y_i \leq r_i^r \\ \dfrac{y_i - r_i^r}{r_i^a - r_i^r}, & r_i^r < y_i < r_i^a \\ \alpha \dfrac{y_i - r_i^a}{r_i^a - r_i^r} + 1, & y_i \geq r_i^a , \end{cases} \tag{4}$$

where α and γ are arbitrarily defined parameters satisfying $0 < \alpha < 1 < \gamma$. Parameter α represents additional increase of the DM's satisfaction over level 1 when a criterion generates outcomes better than the corresponding aspiration level. On the other hand, parameter $\gamma > 1$ represents dissatisfaction connected with outcomes worse than the reservation level.

For outcomes between the reservation and the aspiration levels, the partial achievement function s_i can be interpreted as a membership function μ_i for a fuzzy target. However, such a membership function remains constant with value 1 for all outcomes greater than the corresponding aspiration level, and with value 0 for all outcomes below the reservation level (Fig. 2). Hence, the fuzzy membership function is neither strictly monotonic nor concave thus not representing typical utility for a maximized outcome. The partial achievement function (4) can be viewed as an extension of the fuzzy membership function to a strictly monotonic and concave utility. One may also notice that the aggregation scheme used to build the scalarizing achievement function (2) from the partial ones may also be interpreted as some fuzzy aggregation operator [5]. In other words, maximization of the scalarizing achievement function (2) is consistent with the fuzzy methodology in the case of not attainable aspiration levels and satisfiable all reservation levels while modeling a reasonable utility for any values of aspiration and reservation levels.

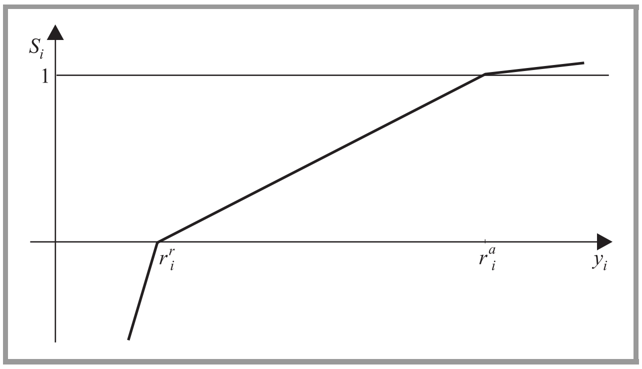


Fig. 2. Partial achievement function (4).

3. Ordered Weighted Averages Refinement of the RPM

The crucial properties of the RPM are related to the max-min aggregation of partial achievements while the regularization is only introduced to guarantee the aggregation monotonicity. Unfortunately, the distribution of achievements may make the max-min criterion partially passive when one specific achievement is relatively very small for all the solutions. Maximization of the worst achievement may then leave all other achievements unoptimized. Nevertheless, the selection is then made according to linear aggregation of the regularization term instead of the max-min aggregation, thus destroying the preference model of the RPM. This can be illustrated with an example of a simple discrete problem of 7 alternative feasible solutions to be selected according to 6 criteria.

Table 1 presents six partial achievements for all the solutions where the partial achievements have been defined according to the aspiration/reservation model (4) thus allocating 1 to outcomes reaching the corresponding aspiration level. All the solutions are efficient. Solution S1 to S5 oversteps the aspiration levels (achievement values 1.2) for four of the first five criteria while failing to reach one of them and the aspiration level for the sixth criterion as well (achievement values 0.3). Solution S6 meets the aspiration levels (achievement values 1.0) for the first five criteria while failing to reach only the aspiration level for the sixth criterion (achievement values 0.3). All the solutions generate the same worst achievement value 0.3 and the final selection of the RPM depends on the total achievement (regularization term). Actually, one of solutions S1 to S5 will be selected as better than S6.

Table 1
Sample achievements with passive max-min criterion

Solutions	a_1	a_2	a_3	a_4	a_5	a_6	min	Σ
S1	0.3	1.2	1.2	1.2	1.2	0.3	0.3	5.4
S2	1.2	0.3	1.2	1.2	1.2	0.3	0.3	5.4
S3	1.2	1.2	0.3	1.2	1.2	0.3	0.3	5.4
S4	1.2	1.2	1.2	0.3	1.2	0.3	0.3	5.4
S5	1.2	1.2	1.2	1.2	0.3	0.3	0.3	5.4
S6	1.0	1.0	1.0	1.0	1.0	0.3	0.3	5.3
S7	0.3	0.3	0.3	1.0	0.6	1.0	0.3	3.5

In order to avoid inconsistencies caused by the regularization, the max-min solution may be regularized according to the ordered averaging rules [13]. This is mathematically formalized as follows. Within the space of achievement vectors we introduce map $\Theta = (\theta_1, \theta_2, \ldots, \theta_m)$ which orders the coordinates of achievements vectors in a nondecreasing order, i.e., $\Theta(a_1, a_2, \ldots, a_m) = (\theta_1(\mathbf{a}), \theta_2(\mathbf{a}), \ldots, \theta_m(\mathbf{a}))$ iff there exists a permutation τ such that $\theta_i(\mathbf{a}) = a_{\tau(i)}$ for all i and $\theta_1(\mathbf{a}) \leq \theta_2(\mathbf{a}) \leq \ldots \leq \theta_m(\mathbf{a})$. The standard max-min aggregation depends on maximization of $\theta_1(\mathbf{a})$ and it ignores values of $\theta_i(\mathbf{a})$ for $i \geq 2$. In order to take into account all the achievement values, one needs to maximize the weighted combination of the ordered achievements thus representing the so-called ordered weighted averaging aggregation [13]. Note that the weights are then assigned to the specific positions within the ordered achievements rather than to the partial achievements themselves. With the OWA aggregation one gets the following RPM model:

$$\max \sum_{i=1}^{m} w_i \theta_i(\mathbf{a}), \tag{5}$$

where $w_1 > w_2 > \ldots > w_m > 0$ are positive and strictly decreasing weights. Actually, they should be significantly decreasing to represent regularization of the max-min order. When differences among weights tend to infinity, the OWA aggregation approximates the leximin ranking of the ordered outcome vectors [14]. Note that the standard RPM model with the scalarizing achievement function (2) can be expressed as the following OWA model [15]:

$$\max \left((1 + \frac{\varepsilon}{m}) \theta_1(\mathbf{a}) + \frac{\varepsilon}{m} \sum_{i=2}^{m} \theta_i(\mathbf{a}) \right).$$

Hence, the standard RPM model exactly represents the OWA aggregation (5) with strictly decreasing weights in the case of $m = 2$ ($w_2 = \varepsilon/2 < w_1 = 1 + \varepsilon/2$). For $m > 2$ it abandons the differences in weighting of the largest achievement, the second largest one, etc., ($w_2 = \ldots = w_m = \varepsilon/m$). The OWA RPM model 5 allows one to distinguish all the weights by introducing increasing series (e.g., geometric ones). One may notice in Table 2 that application of decreasing weights $\mathbf{w} = (0.5, 0.25, 0.15, 0.05, 0.03, 0.02)$ within the OWA RPM enable selection of solution S6 from Table 1.

Table 2
Ordered achievements values

Solutions	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	$A_{\mathbf{w}}$
S1	0.3	0.3	1.2	1.2	1.2	1.2	0.525
S2	0.3	0.3	1.2	1.2	1.2	1.2	0.525
S3	0.3	0.3	1.2	1.2	1.2	1.2	0.525
S4	0.3	0.3	1.2	1.2	1.2	1.2	0.525
S5	0.3	0.3	1.2	1.2	1.2	1.2	0.525
S6	0.3	1.0	1.0	1.0	1.0	1.0	0.650
S7	0.3	0.3	0.3	0.6	1.0	1.0	0.305
w	0.5	0.25	0.15	0.05	0.03	0.02	

An important advantage of the RPM depends on its easy implementation as an expansion of the original multiple criteria model. Actually, even complicated partial achievement functions of the form (4) are strictly increasing and concave, thus allowing for implementation of the entire RPM model (2) by an location problem (LP) expansion [12].

The OWA aggregation is obviously a piecewise linear function since it remains linear within every area of the fixed order of arguments. The ordered achievements used in the OWA aggregation are, in general, hard to implement due to the pointwise ordering. Its optimization can be implemented by expressing in terms of the cumulated ordered achievements $\bar{\theta}_k(\mathbf{a}) = \sum_{i=1}^{k} \theta_i(\mathbf{a})$ expressing, respectively: the worst (smallest) achievement, the total of the two worst achievements, the total of the three worst achievements, etc. Indeed,

$$\sum_{i=1}^{m} w_i \theta_i(\mathbf{a}) = \sum_{i=1}^{m} w_i' \bar{\theta}_i(\mathbf{a}),$$

where $w_i' = w_i - w_{i+1}$ for $i = 1, \ldots, m-1$, and $w_m' = w_m$. This simplifies dramatically the optimization problem since quantities $\bar{\theta}_k(\mathbf{a})$ can be optimized without use of any integer variables [6]. First, let us notice that for any given vector $\mathbf{a}$, the cumulated ordered value $\bar{\theta}_k(\mathbf{a})$ can be found as the optimal value of the following LP problem:

$$\begin{aligned} \bar{\theta}_k(\mathbf{a}) = \min_{u_{ik}} \{ \sum_{i=1}^{m} a_i u_{ik} : \\ \sum_{i=1}^{m} u_{ik} = k, 0 \leq u_{ik} \leq 1 \quad \forall i \}. \end{aligned} \tag{6}$$

The above problem is an LP for a given outcome vector $\mathbf{a}$ while it becomes nonlinear for $\mathbf{a}$ being a vector of variables. This difficulty can be overcome by taking advantage of the LP dual to (6). Introducing dual variable t_k corresponding to the equation $\sum_{i=1}^{m} u_{ik} = k$ and variables d_{ik} corresponding to upper bounds on u_{ik} one gets the following LP dual of problem (6):

$$\begin{aligned} \bar{\theta}_k(\mathbf{a}) = \max_{t_k, d_{ik}} \{ k t_k - \sum_{i=1}^{m} d_{ik} : \\ a_i \geq t_k - d_{ik},\ d_{ik} \geq 0 \quad \forall\ i \}. \end{aligned} \tag{7}$$

Due the duality theory, for any given vector $\mathbf{a}$ the cumulated ordered coefficient $\bar{\theta}_k(\mathbf{a})$ can be found as the optimal value of the above LP problem. It follows from (7) that $\bar{\theta}_k(\mathbf{a}) = \max\ \{k t_k - \sum_{i=1}^{m} (t_k - a_i)_+\}$, where $(.)_+$ denotes the nonnegative part of a number and t_k is an auxiliary (unbounded) variable. The latter, with the necessary adaptation to the minimized outcomes in location problems, is equivalent to the computational formulation of the k–centrum model introduced by [16]. Hence, formula (7) provides an alternative proof of that formulation.

Taking advantages of the LP expression (7) for $\bar{\theta}_i$ the entire OWA aggregation of the partial achievement functions (5) can be expressed in terms of LP. Moreover, in the case of concave piecewise linear partial achievement functions (as typically used in the RPM approaches), the resulting formulation extends the original constraints and criteria with linear inequalities. In particular, for strictly increasing and concave partial achievement functions (4), it can be expressed in the form:

$$\begin{aligned}
&\max \sum_{k=1}^{m} w'_k z_k \\
&\text{s.t.} \\
&z_k = kt_k - \sum_{i=1}^{m} d_{ik} && \forall\, k \\
&\mathbf{x} \in Q,\ y_i = f_i(\mathbf{x}) && \forall\, i \\
&a_i \geq t_k - d_{ik},\ d_{ik} \geq 0 && \forall\, i,k \\
&a_i \leq \gamma(y_i - r_i^r)/(r_i^a - r_i^r) && \forall\, i \\
&a_i \leq (y_i - r_i^r)/(r_i^a - r_i^r) && \forall\, i \\
&a_i \leq \alpha(y_i - r_i^a)/(r_i^a - r_i^r) + 1 && \forall\, i\,.
\end{aligned} \tag{8}$$

4. Weighted OWA Enhancement

Typical RPM model allows weighting of several achievements only by straightforward rescaling of the achievement values [8]. The OWA RPM model enables one to introduce importance weights to affect achievement importance by rescaling accordingly its measure within the distribution of achievements as defined in the so-called weighted OWA aggregation [7], [17]. Let $\mathbf{w} = (w_1, \ldots, w_m)$ and $\mathbf{p} = (p_1, \ldots, p_m)$ be weighting vectors of dimension m such that $w_i \geq 0$ and $p_i \geq 0$ for $i = 1, 2, \ldots, m$ as well as $\sum_{i=1}^{m} p_i = 1$ (typically it is also assumed $\sum_{i=1}^{m} w_i = 1$ but it is not necessary in our applications). The corresponding weighted OWA aggregation of outcomes $\mathbf{a} = (a_1, \ldots, a_m)$ is defined as follows [7]:

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{a}) = \sum_{i=1}^{m} \omega_i \theta_i(\mathbf{a})\,, \tag{9}$$

where the weights ω_i are defined as

$$\omega_i = w^*\big(\sum_{k \leq i} p_{\tau(k)}\big) - w^*\big(\sum_{k < i} p_{\tau(k)}\big)\,, \tag{10}$$

with w^* a monotone increasing function that interpolates points $(\frac{i}{m}, \sum_{k \leq i} w_k)$ together with the point (0.0) and τ representing the ordering permutation for $\mathbf{a}$ (i.e., $a_{\tau(i)} = \theta_i(\mathbf{a})$). Moreover, function w^* is required to be a straight line when the point can be interpolated in this way, thus allowing the WOWA to cover the standard weighted mean with weights p_i as a special case of equal preference weights ($w_i = 1/m$ for $i = 1, 2, \ldots, m$).

Example 1: Consider achievements vectors $\mathbf{a}' = (1,2)$ and $\mathbf{a}'' = (2,1)$. While introducing preferential weights $\mathbf{w} = (0.9, 0.1)$ one may calculate the OWA averages: $A_{\mathbf{w}}(\mathbf{y}') = A_{\mathbf{w}}(\mathbf{y}'') = 0.9 \cdot 1 + 0.1 \cdot 2 = 1.1$. Further, let us introduce importance weights $\mathbf{p} = (0.75, 0.25)$ which means that results under the first achievement are 3 times more important then those related to the second criterion. To take into account the importance weights in the WOWA aggregation (9) we introduce piecewise linear function

$$w^*(\xi) = \begin{cases} 0.9\xi/0.5 & \text{for} \quad 0 \leq \xi \leq 0.5 \\ 0.9 + 0.1(\xi - 0.5)/0.5 & \text{for} \quad 0.5 < \xi \leq 1.0 \end{cases}$$

and calculate weights ω_i according to formula (10) as w^* increments corresponding to importance weights of the ordered outcomes, as illustrated in Fig. 3. In particular, one gets $\omega_1 = w^*(p_1) = 0.95$ and $\omega_2 = 1 - w^*(p_1) = 0.05$ for vector $\mathbf{a}'$ while $\omega_1 = w^*(p_2) = 0.45$ and $\omega_2 = 1 - w^*(p_2) = 0.55$ for vector $\mathbf{a}''$. Finally, $A_{\mathbf{w},\mathbf{p}}(\mathbf{a}') = 0.95 \cdot 1 + 0.05 \cdot 2 = 1.05$ and $A_{\mathbf{w},\mathbf{p}}(\mathbf{a}'') = 0.45 \cdot 1 + 0.55 \cdot 2 = 1.55$. Note that one may alternatively compute the WOWA values by using the importance weights to replicate corresponding achievements and calculate then OWA aggregations. In the case of our importance weights $\mathbf{p}$ we need to consider three copies of achievement a_1 and one copy of achievement a_2

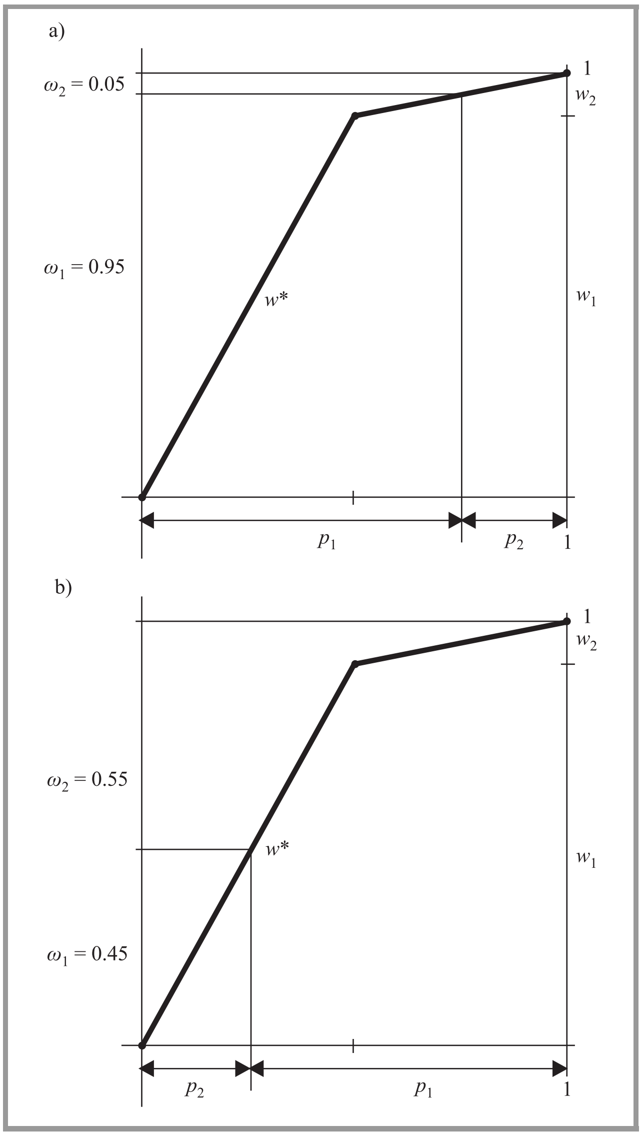


Fig. 3. Definition of weights ω_i for Example 1: (a) vector $\mathbf{a}' = (1,2)$; (b) vector $\mathbf{a}'' = (2,1)$.

thus generating vectors $\tilde{\mathbf{a}}' = (1,1,1,2)$ and $\tilde{\mathbf{a}}'' = (2,2,2,1)$ of four equally important achievements. Original preferential weights must be then applied respectively to the average of the two smallest outcomes and the average of two largest outcomes. Indeed, we get $A_{\mathbf{w},\mathbf{p}}(\mathbf{a}') = 0.9 \cdot 1 + 0.1 \cdot 1.5 = 1.05$ and $A_{\mathbf{w},\mathbf{p}}(\mathbf{a}'') = 0.9 \cdot 1.5 + 0.1 \cdot 2 = 1.55$. We will further formalize this approach and take its advantages to build LP computational models. □

The WOWA may be expressed with more direct formula where preferential (OWA) weights w_i are applied to averages of the corresponding portions of ordered achievements (quantile intervals) according to the distribution defined by importance weights p_i [18], [19]:

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{a}) = \sum_{i=1}^{m} w_i m \int_{\frac{i-1}{m}}^{\frac{i}{m}} F_{\mathbf{a}}^{(-1)}(\xi)\, d\xi\,, \tag{11}$$

where $F_{\mathbf{a}}^{(-1)}$ is the stepwise function $F_{\mathbf{a}}^{(-1)}(\xi) = \theta_i(\mathbf{a})$ for $i - 1/m < \xi \le i/m$. It can also be mathematically formalized as follows. First, we introduce the right-continuous cumulative distribution function (cdf):

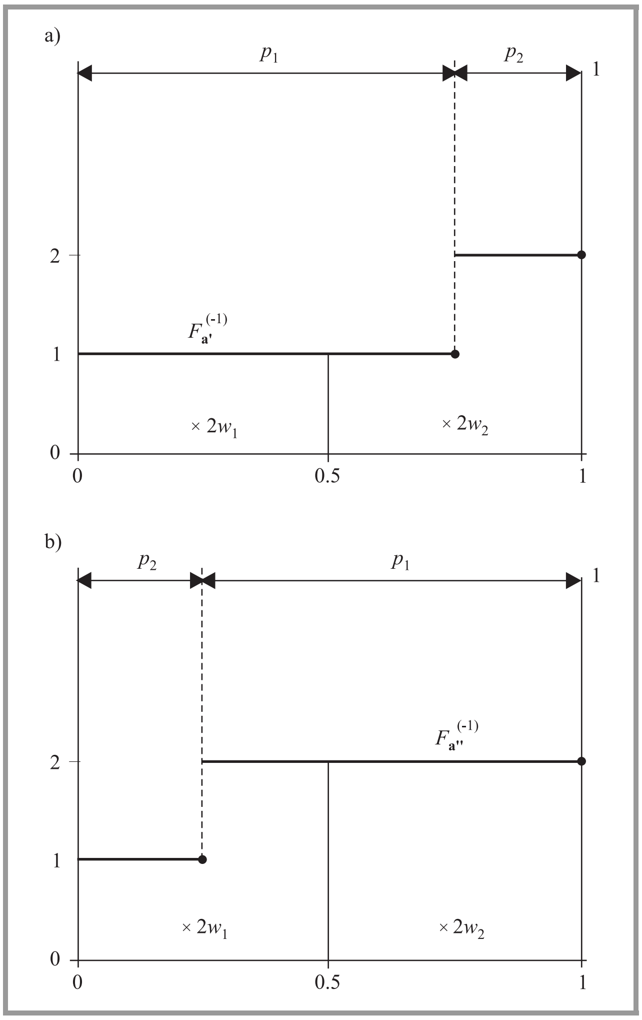


Fig. 4. Formula (11) applied to calculations in Example 1: (a) vector $\mathbf{a}' = (1,2)$; (b) vector $\mathbf{a}'' = (2,1)$.

$$F_{\mathbf{a}}(d) = \sum_{i=1}^{m} p_i \delta_i(d)\,, \tag{12}$$

where $\delta_i(d) = 1$ if $a_i \le d$ and 0 otherwise. Next, we introduce the quantile function $F_{\mathbf{a}}^{(-1)}$ as the left-continuous inverse of the cumulative distribution function $F_{\mathbf{a}}$, ie., $F_{\mathbf{a}}^{(-1)}(\xi) = \inf\{\eta : F_{\mathbf{a}}(\eta) \ge \xi\}$for $0 < \xi \le 1$. Figure 4 illustrates application of formula (11) for computation of the WOWA aggregations in Example 1.

Let us recall the RPM applied to the example of seven alternatives as given in Table 1. For instance applying importance weighting $\mathbf{p} = (\frac{4}{12}, \frac{3}{12}, \frac{2}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$ to solution achievements from Table 1 and using them together with the OWA weights $\mathbf{w}$ from Table 2 one get the WOWA aggregations from Table 3. The corresponding RPM method selects than solution S6, similarly to the case of equal importance weights. On the other hand, when increasing the importance of the last outcome achievements with $\mathbf{p} = (\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{7}{12})$ one get the WOWA values from Table 4.

Formula (11) defines the WOWA value applying preferential weights w_i to importance weighted averages within quantile intervals. It may be reformulated with the tail averages:

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{a}) = \sum_{k=1}^{m} w'_k m L\Big(\mathbf{a}, \mathbf{p}, \frac{k}{m}\Big)\,, \tag{13}$$

where $L(\mathbf{y}, \mathbf{p}, \xi)$ is defined by left-tail integrating of $F_{\mathbf{y}}^{(-1)}$, i.e.,

$$L(\mathbf{y}, \mathbf{p}, \xi) = \int_0^{\xi} F_{\mathbf{y}}^{(-1)}(\alpha) d\alpha \tag{14}$$

and weights $w'_k = w_k - w_{k+1}$ for $k = 1, \ldots, m-1$ and $w'_m = w_m$.

Graphs of functions $L(\mathbf{a}, \mathbf{p}, \xi)$ (with respect to ξ) take the form of convex piecewise linear curves, the so-called absolute Lorenz curves [20] connected to the relation of the second order stochastic dominance (SSD). Therefore, formula (13) relates the WOWA average to the SSD consistent risk measures based on the tail means provided that the importance weights are treated as scenario probabilities.

According to (14), values of function $L(\mathbf{a}, \mathbf{p}, \xi)$ for any $0 \le \xi \le 1$ can be given by optimization:

$$L(\mathbf{a}, \mathbf{p}, \xi) = \min_{s_i} \Big\{ \sum_{i=1}^{m} a_i s_i : \sum_{i=1}^{m} s_i = \xi, \quad 0 \le s_i \le p_i \quad \forall\, i \Big\}. \tag{15}$$

Introducing dual variable t corresponding to the equation $\sum_{i=1}^{m} s_i = \xi$ and variables d_i corresponding to upper

Table 3
WOWA RPM selection with importance weights $\mathbf{p} = (\frac{4}{12}, \frac{3}{12}, \frac{2}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$

w	0.5		0.25		0.15		0.05		0.03		0.02		$A_{\mathbf{w},\mathbf{p}}(\mathbf{a})$
S1	0.3	0.3	0.3	0.3	0.3	1.2	1.2	1.2	1.2	1.2	1.2	1.2	0.4575
S2	0.3	0.3	0.3	0.3	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	0.525
S3	0.3	0.3	0.3	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	0.6375
S4	0.3	0.3	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	0.75
S5	0.3	0.3	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	0.75
S6	0.3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.825
S7	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.6	1.0	1.0	0.3185

Table 4
WOWA RPM selection with importance weights $\mathbf{p} = (\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{7}{12})$

w	0.5		0.25		0.15		0.05		0.03		0.02		$A_{\mathbf{w},\mathbf{p}}(\mathbf{a})$
S1	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.2	1.2	1.2	1.2	0.345
S2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.2	1.2	1.2	1.2	0.345
S3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.2	1.2	1.2	1.2	0.345
S4	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.2	1.2	1.2	1.2	0.345
S5	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.2	1.2	1.2	1.2	0.345
S6	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.0	1.0	1.0	1.0	1.0	0.3525
S7	0.3	0.3	0.3	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5125

bounds on s_i one gets the following LP dual expression of $L(\mathbf{a},\mathbf{p},\xi)$:

$$L(\mathbf{a},\mathbf{p},\xi) = \max_{t,d_i} \Big\{ \xi t - \sum_{i=1}^{m} p_i d_i : \quad t - d_i \le a_i,\ d_i \ge 0 \quad \forall\, i \Big\}. \qquad (16)$$

Following (13) and (16) taking into account piecewise linear partial achievement functions (4) one gets finally the following model for the WOWA reference point method with piecewise linear partial achievement functions (4):

$$\begin{aligned}
&\max \sum_{k=1}^{m} w'_k z_k \\
&\text{s.t.} \\
&z_k = k t_k - m \sum_{i=1}^{m} p_i d_{ik} && \forall\, k \\
&\mathbf{x} \in Q,\ y_i = f_i(\mathbf{x}) && \forall\, i \qquad (17)\\
&a_i \ge t_k - d_{ik},\ d_{ik} \ge 0 && \forall\, i,k \\
&a_i \le \gamma(y_i - r_i^r)/(r_i^a - r_i^r) && \forall\, i \\
&a_i \le (y_i - r_i^r)/(r_i^a - r_i^r) && \forall\, i \\
&a_i \le \alpha(y_i - r_i^a)/(r_i^a - r_i^r) + 1 && \forall\, i\,.
\end{aligned}$$

5. Illustrative Example

In order to illustrate the WOWA RPM performances let us analyze the multicriteria problem of information system selection. We consider a billing system selection for a telecommunication company. The decision is based on 6 criteria related to the system reliability, processing efficiency, investment costs, installation time, operational costs, and warranty period. All these attributes may be viewed as criteria, either maximized or minimized.

Table 5 presents all the criteria with their measures units and optimization directions. There are also set the aspiration and reservation levels for each criterion as well as the importance weights (not normalized) for several achievements. Five candidate billing systems have been accepted for the final selection procedure. All they meet the minimal requirements defined by the reservation levels.

Table 6 shows for all the systems (columns) their criteria values y_i and the corresponding partial achievement values a_i. The latter are computed according to the piece-wise linear formula (4) with $\alpha = 0.1$.

Table 7 presents for all the systems (columns) their partial achievement values ordered from the worst to the best taking into account replications according to the importance weights allowing for easy WOWA aggregation computations following formula (11). One may notice that except of system D all the other systems have the same worst achievement value $\min_i a_i = 0.33$. Selection among systems A, B, C and E depends only on the regularization of achievements aggregation used in the RPM approach. The WOWA RPM method taking into account the importance weights together with the preferential weights $\mathbf{w} = (0.6, 0.2, 0.1, 0.05, 0.03, 0.02)$ points out system C

Table 5
Criteria and their reference levels for the sample billing system selection

Criteria	f_1 reliability	f_2 efficiency	f_3 invest. cost	f_4 install. time	f_5 oprnl. cost	f_6 warranty period
Units	1–10	CAPS	mln PLN	months	mln PLN	years
Optimization	max	max	min	min	min	max
Reservation	8	50	2	12	1.25	0.5
Aspiration	10	200	0	6	0.5	2
Importance weights	0.3	0.3	0.1	0.1	0.3	0.1

Table 6
Criteria values y_i and individual achievements a_i for five billing systems

i	System A		System B		System C		System D		System E	
	y_i	a_i	y_i	a_i	y_i	a_i	y_i	a_i	y_i	a_i
1	10	1.00	9	0.50	10	1.00	9	0.50	10	1.00
2	200	1.00	100	0.33	170	0.80	90	0.27	150	0.67
3	1	0.50	0.3	0.85	0.8	0.60	0.2	0.90	0.5	0.75
4	8	0.67	3	1.05	8	0.67	8	0.67	5	1.02
5	1	0.33	1	0.33	0.6	0.87	0.2	1.04	1	0.33
6	2	1.00	2	1.00	1	0.33	2	1.00	1.5	0.67
$\min a_i$		0.33		0.33		0.33		0.27		0.33
$\sum a_i$		4.50		4.06		4.27		4.38		4.44

as the best one. These selection cannot be done if using the classical RPM with regularization based on the average achievements. Actually, the standard RPM (2) will select system A as better than all the others. Certainly, the WOWA RPM selection will change dramatically when decreasing importance of criterion f_5 and increasing importance of f_6.

Table 7
WOWA RPM selection for five billing systems

w	A	B	C	D	E
0.6	0.33	0.33	0.33	0.27	0.33
	0.33	0.33	0.60	0.27	0.33
0.2	0.33	0.33	0.67	0.27	0.33
	0.50	0.33	0.80	0.50	0.67
0.1	0.67	0.33	0.80	0.50	0.67
	1.00	0.33	0.80	0.50	0.67
0.05	1.00	0.50	0.87	0.67	0.67
	1.00	0.50	0.87	0.90	0.75
0.03	1.00	0.50	0.87	1.00	1.00
	1.00	0.85	1.00	1.04	1.00
0.02	1.00	1.00	1.00	1.04	1.00
	1.00	1.05	1.00	1.04	1.02
$A_{\mathbf{w},\mathbf{p}}$	0.47	0.37	0.56	0.37	0.45

6. Conclusions

The reference point method is a very convenient technique for interactive analysis of the multiple criteria optimization problems. It provides the DM with a tool for an open analysis of the efficient frontier. The interactive analysis is navigated with the commonly accepted control parameters expressing reference levels for the individual objective functions. The partial achievement functions quantify the DM satisfaction from the individual outcomes with respect to the given reference levels. The final scalarizing function is built as the augmented max-min aggregation of partial achievements which means that the worst individual achievement is essentially maximized but the optimization process is additionally regularized with the term representing the average achievement. The regularization by the average achievement is easily implementable but it may disturb the basic max-min aggregation. In order to avoid inconsistencies caused by the regularization, the max-min solution may be regularized by taking into account also the second worst achievement, the third worse and so on, thus resulting in much better modeling of the reference levels concept [21].

The OWA aggregation with monotonic weights combines all the partial achievements allocating the largest weight to the worst achievement, the second largest weight to the second worst achievement, the third largest weight to the third worst achievement, and so on. It approximates nucleolar RPM introducing explicit scalarizing achievement function to be interpreted as utility. Further following the concept of weighted OWA [7] the importance weighting of several achievements may be incorporated into the RPM. Such a WOWA enhancement of the RPM uses importance weights to affect achievement importance by rescaling accordingly its measure within the distribution of achievements rather than straightforward rescaling of achievement values [8]. The ordered regularizations are more complicated in implementation due to the requirement of pointwise ordering of partial achievements. However, the recent progress in optimization methods for ordered averages [6] allows one to implement the OWA RPM quite effectively by taking advantages of piecewise linear expression of the cumulated ordered achievements. Similar, model can be achieved for the WOWA RPM. Actually, in the case of concave piecewise linear partial achievement functions (typically used in the RPM), the resulting formulation extends the original constraints and criteria with simple linear inequalities thus allowing for a quite efficient implementation.

Acknowledgements

The research was partially supported by the Polish Ministry of Science and Higher Education under the research grant N N516 4307 33.

References

[1] A. P. Wierzbicki, “A mathematical basis for satisficing decision making”, *Math. Model.*, vol. 3, pp. 391–405, 1982.

[2] J. Granat and M. Makowski, “ISAAP – interactive specification and analysis of aspiration-based preferences”, *Eur. J. Opnl. Res.*, vol. 122, pp. 469–485, 2000.

[3] A. Lewandowski and A. P. Wierzbicki, *Aspiration Based Decision Support Systems – Theory, Software and Applications*. Berlin: Springer, 1989.

[4] W. Ogryczak and S. Lahoda, “Aspiration/reservation decision support – a step beyond goal programming”, *J. MCDA*, vol. 1, pp. 101–117, 1992.

[5] A. P. Wierzbicki, M. Makowski, and J. Wessels, *Model Based Decision Support Methodology with Environmental Applications*. Dordrecht: Kluwer, 2000.

[6] W. Ogryczak and T. Śliwiński, “On solving linear programs with the ordered weighted averaging objective”, *Eur. J. Opnl. Res.*, vol. 148, pp. 80–91, 2003.

[7] V. Torra, “The weighted OWA operator”, *Int. J. Intell. Syst.*, vol. 12, pp. 153–166, 1997.

[8] F. Ruiz, M. Luque, and J. M. Cabello, “A classification of the weighting schemes in reference point procedures formultiobjective programming”, *J. Opnl. Res. Soc.*, 2008 (forthcoming), doi: 10.1057/palgrave.jors.2602577.

[9] A. P. Wierzbicki, “On completeness and constructiveness of parametric characterizations to vector optimization problems”, *OR Spectrum*, vol. 8, pp. 73–87, 1986.

[10] K. Miettinen and M. M. Mäkelä, “On scalarizing functions in multiobjective optimization”, *OR Spectrum*, vol. 24, pp. 193–213, 2002.

[11] W. Ogryczak, “Preemptive reference point method”, in *Multicriteria Analysis — Proceedings of the XIth International Conference on MCDM*, J. Climaco, Ed. Berlin: Springer, 1997, pp. 156–167.

[12] W. Ogryczak, K. Studziński, and K. Zorychta, “DINAS: a computer-assisted analysis system for multiobjective transshipment problems with facility location”, *Comp. Opns. Res.*, vol. 19, pp. 637–647, 1992.

[13] R. R. Yager, “On ordered weighted averaging aggregation operators in multicriteria decision making”, *IEEE Trans. Syst., Man Cyber.*, vol. 18, pp. 183–190, 1988.

[14] R. R. Yager, “On the analytic representation of the Leximin ordering and its application to flexible constraint propagation”, *Eur. J. Opnl. Res.*, vol. 102, pp. 176–192, 1997.

[15] W. Ogryczak, “On goal programming formulations of the reference point method”, *J. Opnl. Res. Soc.*, vol. 52, pp. 691–698, 2001.

[16] W. Ogryczak and A. Tamir, “Minimizing the sum of the *k* largest functions in linear time”, *Inform. Proc. Let.*, vol. 85, pp. 117–122, 2003.

[17] X. Liu, “Some properties of the weighted OWA operator”, *IEEE Trans. Syst. Man Cyber. B*, vol. 368, pp. 118–127, 2006.

[18] W. Ogryczak and T. Śliwiński, “On optimization of the importance weighted OWA aggregation of multiple criteria”, in *International Conference Computational Science and Its Applications ICCSA 2007*, *LNCS*, vol. 4705. Heidelberg: Springer, 2007, pp. 804–817.

[19] W. Ogryczak and T. Śliwiński, “On decision support under risk by the WOWA optimization”, in *Ninth European Conference on Symbolic and Quanlitative Approaches to Reasoning with Uncertainty ECSQARU 2007*, *LNAI*, vol. 4724. Heidelberg: Springer, 2007, pp. 779–790.

[20] W. Ogryczak and A. Ruszczyński, “Dual stochastic dominance and related mean-risk models”, *SIAM J. Opt.*, vol. 13, pp. 60–78, 2002.

[21] M. M. Kostreva, W. Ogryczak, and A. Wierzbicki, “Equitable aggregations and multiple criteria analysis”, *Eur. J. Opnl. Res.*, vol. 158, pp. 362–367, 2004.

Włodzimierz Ogryczak is a Professor and Deputy Director for Research 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 models, 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.
e-mail: wogrycza@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

On Subjective Trust Management

Michał Majdan

Abstract—Trust and reputation management is gaining nowadays more attention then ever as online commodity exchange and other open virtual societies became a widespread reality. Most widely used computational models use reputation metrics as global property assigned to each party. More sophisticated models try to use reputation as subjective property. While introducing subjective reputation there arise a need to model preferences of agents. In this paper we propose to use weighted ordered weighted average (WOWA) operator to support the decision maker in assessing available evidence about other's party behavior. The WOWA aggregation is defined by two weighting vectors: the preferential weights assigned to the ordered quantities and the importance weights assigned to several attributes. It allows one to express both the preference regarding sources of information by the corresponding importance weights and the compensation between attribute values aggregated by the preferential weights.

Keywords— decision support, management concepts, multicriteria reputation model.

1. Introduction

Internet society has become reality. Nowadays more and more of commodities are exchanged via online commerce sites. Success of online auctions like e-Bay or Allegro has shown how dramatically fast more and more aspects of human life are moving online. But commerce of goods and services is not the only domain changed by the Internet. Peer-to-peer (p2p) networks exchanging software and data are getting more and more attention too. Wikipedia, an online encyclopedia constantly developed by anonymous users sharing their wisdom end expertise with others. We rely more and more on people or agents whom we probably are never going to meet face to face in "real" life, probably we are not going to establish long term relationship with most of them. Therefore old ways of assessing transaction counterparts are of little use in online world. Thus users try to assure security for online life by designing and implementing trust management systems, algorithms and protocols that are to provide trust between members of such systems.

There are various definitions of trust and there are differences among researchers on how define trust. Most popular in literature is definition by D. Gambetta [1]:

"*Trust is a particular level of subjective probalility with which an agent will perform a particular action both before he can monitor such action (or independently of his capacity of ever to be able to monitor it) and in context in which it affects own action*".

The "probability" mentioned above is "subjective" since according to this definition the same agent may be perceived differently by other members of the same virtual community. Not all models follow this assumption some aim to compute some kind of global evaluation of a certain agent (it will be mentioned below).

The second component that is very tightly coupled to the concept of trust is reputation. Generally reputation is "what-is-said" about certain agent in a given community. That's why if we consider any subset of this community the reputation of the same agent may differ. It's important to notice that reputation is not what a certain agent "thinks" about the other, reputation is a kind of aggregation of opinions communicated by a given group of agents. This concept coming from the cognitive theory of reputation is exploited by J. Sabater *et al.* in [2].

2. Trust and Reputation Models

Computational trust and reputation management models usually use 4 sources of information to build view on a particular agent:

1. Own direct experience – these are valuations that come from agent's own interactions with the target agent. The credibility of those valuations is perfect, so they are most valuable for any trust model. The problem is that in real world situations rarely there exists more than 1 interaction that involves the same 2 agents. So the ability to collect enough data based on direct experience is very limited.

2. Observations – in some environments agents are able to observe performance of the others. Like with the direct experience those evaluations are highly valuable. Very rarely though open virtual communities offer the ability to monitor interactions of others to everybody.

3. Witness information – these are valuations provided by other agents regarding third party. Those can be vulnerable to untrue valuations malicious agents who are willing to destroy others reputations or contrary inadequately promote others. Apart from those threats different agents assess others performance differently. Behavior totaly inappropriate in the eyes of one can be correct when assessed by the other. Witness information is actually something that constitutes reputation as word-of-mouth.

4. Context information – each interaction is performed in a particular context. Context can be described as a vector of parameters different for each system. For online auction example this can be the value of the transaction, whereabouts of the seller or buyer, means of payment, type of merchandise.

Trust models usually recognize at least 2 roles of an agent that could be considered when determining trust metrics. These are firstly trust on whether particular agent will perform as we expect him to do and secondly whether he can be trusted as an informant. This can be further divided into roles based on what kind of activity we require from the particular agent: for example, one can be trusted seller but fail at delivering services. Other important dimension that is used in trust and reputation models is time. Usually trust is a function of reputation and the time elapsed from the events that update the reputation. This reflects belief that conclusions based on past behavior are less accurate the further in the past we look. What differs among various reputation based models is what kind of so called third party information is exchanged.

Some of the models assume that peers exchange their overall evaluation of "other party" some call it image [2]. In such situation whole evaluation process as well as the criteria may be totally different at each peer. But clear advantage of such models is that the information traffic is minimal.

Second kind of reputation based models assume that agents exchange evaluations on transactions they had with the other party like in e-Bay feedback system. This allows to incorporate time into the model as receiver of such information is able to for example disregard evaluations associated with transactions older then arbitrary threshold. Such models are most widely used in online commerce sites, for example e-Bay, we will look into this later. Still such models assume that criteria each evaluator uses and his utility is not known and not shared among members of such virtual community.

Another thing that needs to be considered while designing trust model is source reliability. As it was stated above, most of the information that particular agents has to rely on is information provided by third party. It's crucial to eliminate information coming from unreliable sources. The source may be lying on purpose in order to increase own reputation or reputation of some associated identity or to decrease others party reputation. Models developed so far address this problem various ways. One way is by introducing pre-trusted peers [3] and forcing agents to put at least some amount of "trust" in the pre-trusted peers. Others set up protocol to detect malicious peers by asking them about things already known to the asking party. If answer is different such peer's reliability as an informant is decreased. This approach is used in models that assume that the same transaction may be observed by more then 2 involved parties.

One other thing that can be considered in trust/reputation models is agent's confidence about his judgment. When asked about opinion on some other party agent can provide his rating noting that he is sure about this opinion only up to the certain level. This means that each rating is a vector of two values. This is especially applicable to models using reputation as a message exchanged between parties. In such case agent can express using confidence the fact that he had very little experience with the assessed party. Models using transaction rating do not require this additional information. Confidence value associated with specific valuation make the aggregation procedure significantly more complex. Up to now quite large number of trust models have been developed. Most important are presented below. We have focused on how user preferences are incorporated in those models and how valuations are aggregated.

3. Selected Trust Management Concepts Review

3.1. Online Auctions Reputation Models

The most widely used are very simple reputation based systems like e-Bay feedback system. They are based on three valued feedback provided by transaction parties on how they assess the transaction. Positive values mean that the party is satisfied with partner's performance. Neutral values usually mean that transaction is judged as sufficiently correct. Negative feedback values usually mean that transaction was highly unacceptable. It turns out that vast majority of feedback is positive with very little neutral or negative feedback. But there exists significant percent of transactions with no feedback at all. The problem is that giving negative or neutral feedback threatens that other party will retaliate, so people who actually dislike the performance of their transaction partner choose not to provide feedback instead of giving negative. Transaction evaluations (–1,0,1) are summarized over certain period of time (e-Bay – 6 months). Although simple and vulnerable to false information the above mechanism proved to be usable since online auction and general open e-Commerce sites are successful.

Preference structure representation. User preferences are not expressed directly but through feedback value. It's not known what made agent to provide certain value.

3.2. Probabilistic Models

The eigen-trust [3] algorithm has been proposed as a mean to provide trust in the anonymous peer-to-peer networks. The measure of trust is a normalized value between 0 an 1 reflecting number of satisfactory and unsatisfactory transactions with a given peer versus all peers. Since not all peers interact with each other the experiences of others are the only source of information. The party gathering information may ask his direct counterparts about their opinion on a specific target. Their opinions are weighted by the trust value that given peer places in them. Doing further this way one party may ask his friends about their opinions. This is done by multiplying consecutive local trust values. Such computations converge under certain assumptions to the global trust vector that shows how much trust the system as a while places in its members.

The eigen-trust model uses concept of transitive trust. Valuations exchanged between peers refer to the peers reputation not individual transactions. It is assumed that the preference structure of each peer is the same, namely it's the difference between successful interactions and unsuccessful ones. The only piece of subjectivity is when agent is to evaluate whether certain interaction was successful or not. Another example of probabilistic approaches is model proposed by Shillo *et al.* [4]. In this model each interaction is assessed as either good or bad. The personal trust value that one agent places into another is $Q = e/n$, where n is a number of observed interactions and e is a number of positive ones (when other party behaved honestly). An agent can ask other agents about their impressions on observed interactions. Model assumes that agents do not lie about interactions they observe therefore while considering messages from others, there is no need to resolve conflicts (the same interactions assessed differently by different agents).

3.3. Abdul-Rachman, Heils Model

Some models use linguistic labels instead of numerical ratings to represent social valuations. This model is especially interesting as it tries to deal with the issue of different perception of the same information by different agents (people). The idea of this model is that assessing agent can give the other party one of 4 labels based on his opinion about the other. These labels are namely:

- very trustworthy (VT),
- trustworthy (T),
- untrustworthy (U),
- very untrustworthy (VU).

Agent maintains number of interactions of each of the above categories for every other agent in the society (system). The general assessment of the agent is label with maximum number of interactions that "support" this label. In case of two or more labels having the same count model introduces uncertaintly measures into assessment.

The most interesting part of the model specification is how it deals with the reports (opinions) provided by other members of the system. If an agent receives information from the other agent assessing some third party as for example VT and the agent has previously assessed the same other party as only T he will adjust any future opinions received from the same agent accordingly, that is he will lower them. The preference structure at this model is not externalized by the agents but it is expressed indirectly by comparing assessments of the same information.

3.4. OWA Trust

In 1988 Yager [5] proposed new aggregation operator ordered weighted average (OWA). It is similar to weighted mean operator. Weights though are not assigned to particular criteria but to the values within criteria permutation from the biggest value to the smallest value. Formal definition of the operator is as follows: given vector of n criteria we have x_i for $i = 1 \ldots n$ and preferential weights vector w_i for $i = 1 \ldots n$ while $\sum_{i=1}^{n} w_i = 1$. The OWA operator is defined as follows:

$$\sum_{i=1}^{n} w_i x_{\sigma(i)} \, ,$$

where $\sigma(i)$ is certain permutation of vector x that

$$x_{\sigma(1)} \geq x_{\sigma(2)} \geq x_{\sigma(3)} \geq \ldots \geq x_{\sigma(n)} .$$

This operator is used in a trust management model proposed in [6]. Local reputation values are calculated as $t_{ij} = 1$ or $t_{ij} = 0$ if the given peer would have only one successful or unsuccessful interaction with the other peer. If there was more interaction between two peers the overall local reputation is calculated as follows:

$$r_{ij}^{n} = \alpha^{n} r_{ij}^{n-1} + (1 - \alpha^{n}) t_{ij}^{n} .$$

The α^n parameter is accounts for aggregation "freshness", that is how past transactions are important compared with the last one. High values are for the case when past values are far more important than the last one otherwise last interaction is more important. Local reputation is of little use if interactions between same peer pairs are rear. This is almost always the case in online society systems. For such situation model provides the concept of network reputation. Local reputation values from the pool of voting peers are aggregated using the OWA technique. Local reputation values are ordered and aggregated using the set of weights. Authors do not impose any procedure on selecting weights for a given model realization. Authors show that selection of weights reflects decision maker's attitude towards situation.

4. Multicriteria Reputation Model

The model presented in this paper is addressing the potential problem associated with reputation representation. As it was mentioned above models exchange either information about other agents or about transactions. In both cases they exchange aggregated social evaluations build on how they personally perceive other users or certain transactions. Such approach hide the preference structures of the agents. Messages biased by the ones preference structure may lead to wrong assessment of other agents. The proposition is to make information exchanged between agents as objective as possible and to allow consumers of this information to make their own judgments. Models that assume exchange of social valuations of other agents bias the evaluations more than ones exchanging data on transactions so in the presented concept the content of the message will be associated with the single transaction.

4.1. Outline

The first thing to do implementing the model is to set up a set of criteria associated with the single transaction.

Next we need a common way to express the preference structure of each member of the system. This need to be as easy as possible and selected a priori since in a general case agents may not be able to modify their preference structures assessing each interaction.

We need to set up an algorithm to aggregate interactions and make a decision based on the criteria whether to trust or not to trust.

4.2. Selecting Criteria

Setting up a group of criteria is the most vulnerable part of the model. It is very tightly coupled with the actual domain the model being implemented into. The important part is that calculating satisfaction levels for this criteria should be as objective as possible. If it is not we are threatened to exchange single biased social valuation into a set of them which is far less attractive but still can provide some advantage. We can consider following example criteria:

- latency of payment or delivery in case of an auction service;
- price (compared to other items of the same kind in the service in the same time);
- number of e-mail's exchanged between buyer and seller;
- response time in case of a question;
- download speed in case of p2p network;
- number of errors during download, etc.

For successful application of aggregation operator, criteria satisfaction measures need to be commeasurable, therefore they need to be normalized. Since selection of criteria is out of the scope of this paper this topic won't be elaborated further.

4.3. Setting up Preference Structure

The preference structure implemented in the presented model is twofold. First setup is associated with the issue of aggregating information about past transactions reported by other users. This is actually aggregating reputation metric with regard to each of the criteria separately. Every trust management model based on reputation has its own way of aggregating reputation data. This aspect is discussed in detail in [7]. Previous section gives short overview of this topic with regard to the selected models described above. In this paper author proposes one possible aggregation technique with very significant expression power.

An important aspect to consider is source reliability. For now we do not go far into this area. As it was mentioned in the beginning there are being developed various algorithms and protocols that deal with this problem. For now we assume that given agent is able to assess the reliability of the information coming from other agents, and eliminates messages from malicious informants. It's worth to notice that being a malicious informant may not come with failing with delivering as expected. Presented model is to deal with assessing trust with regard to the agent performance in delivering merchandise, data, payments ... and not reputation valuations (criteria satisfaction levels) associated with other members of the community.

The next aspect addressed by trust management models is information being biased by informants own attitude towards assessed situations. This is the problem that presented model is trying to answer. As mentioned above satisfaction levels of the selected criteria should be objectively calculated possibly by the system itself, or normalized in order to establish common view of this values by all agents. The next problem is how agent may express his preferences in order to judge whether he can trust the other or not. This involves comparing calculated trust value with arbitrary threshold (independent for each agent) in order to make a decision to go on with the interaction or to retreat. Setting up preference structure should be as easy and understandable but at the same time needs to have as much expressive power as possible to cover possibly wide range of preferences.

4.4. Weighted Ordered Weighted Average

Weighted ordered weighted average (WOWA) aggregation operator was proposed by V. Torra [8] as a generalization to previous ordered weighted operator proposed by Yager [5]. Torras conception is based on two weighting vectors:

- preferential weights vector w_i[1], associated with values permutation from the highest value to the lowest;
- importance weights vector p_i associated with each of the aggregated criteria.

Formal WOWA definition is described below:

$$WOWA(x_1, \dots x_n) = \sum_{i=1}^{n} \omega_i x_i,$$

while weights ω are constructed as follows:

$$\omega_i = w^*(\sum_{j \le i} p_{\sigma(j)}) - w^*(\sum_{j < i} p_{\sigma(j)}).$$

Function w^* interpolates points $(i/n, \sum_{j \le i} w_j)$ and point (0,0) if points can be interpolated using linear functions they should be interpolated in this way. In case when preferential weights p_i are equal and sum up to one, WOWA becomes standard OWA operator with preferential weights w_i. When preferential weights are equal and sum up to one, WOWA operator becomes weighted average operator. The WOWA aggregation generalizes both OWA and weights average and its actual value are always somewhere in between those aggregation methods.

4.5. Aggregation of Interactions

We need to consider how information regarding past interaction involving assessed agent is being aggregated. We

[1] In general case number of preferential weights can be higher than the number of values aggregated [9].

should aggregate values for each criterion to present to user single vector of aggregated criteria satisfaction levels. We propose to consider use of the WOWA aggregation technique as well suited approach for this case. As mentioned above particular agent while assessing reliability of the informer can use several techniques. However the problem itself is not an easy task and very often agents cannot a priori assess the reliability of others, especially in environments when very rarely there exists more than one interaction involving particular pair of agents. If environment/system allows to monitor reliability of the information, agent can derive a kind of rating for messages he acquired from other parties. If no such functionality is available, messages can be arranged with respect to the date they were created.

In either case derived ratings can be used to establish the vector of importance weights as described in WOWA definition. WOWA gives us also possibility to express attitude towards values of selected criteria satisfaction levels. If particular agent requires that all messages regarding interactions with other party should show high level of performance with respect of a given criteria, he should set preferential weights to form an “anding” operator. If he requires at least one message to be highly satisfactory he should form more “oring” operator. The question of “orness” and “andness” are described in [5].

4.6. Aggregation of Criteria

At this point decision maker is presented with a set of aggregated criteria satisfaction levels. This actually sets up a multicriteria analysis situation. If the situation would be to choose counterpart that can be most trusted with regard to this situation the whole multicriteria analysis methodology and tools could be used to support decision maker (DM) at choosing the best alternative. For instance in case of human DM the reference point method could be successfully applied. But in most trust management situations there exists only a pair of agents and for each of them the interaction requires binary go/not go decision. Further in many cases agent (DM) is not human and cannot modify its preferences interactively. Rather it can be equipped with some a priori set preference structure to filter the available data and make decisions comparing calculated metrics with arbitrary selected threshold. Again WOWA can be used as a scalarizing function. Agent, for example user of online auction service, can establish importance weights to express tradeoffs between criteria. The preferential weights in this are case also used to express how many criteria in general have to be satisfied and to what extent. Calculated overall value is then compared against individual threshold and decision to go with the interaction is based upon this.

5. Conclusions and Further Work

This paper presents outline of trust management model that puts stress on subjective interpretation of information in decision making process. It shows possible way of setting up preference structure by parameters of WOWA aggregation operator. Model requires that criteria selected to describe single interaction can be objectively measured and subjectively assessed. While constructing vector of criteria satisfaction levels it has to be assured that its components are commeasurable.

There are still issues that have to be elaborated. Objective selection and measurement of criteria much depends on particular system and environment but is crucial for successful application of above technique. Some work needs to be done to check real life scenarios if such criteria can be monitored with satisfactory level of objectivity. In case of interaction aggregation setting up importance weights is much dependent on agents reliability verification method. If such method is nonexistent or can be compromised easily it might be good to consider setting all importance weights to equal value reducing WOWA to OWA operator.

It’s planned to apply this model to a real life situation of online auction service. If all agents are forced to express their preferences in a common and understandable form there arises opportunity to analyze users of such system with regard to their preference structure and to adopt system accordingly.

Acknowledgement

This paper has been partially founded by the Polish Ministry of Science and Higher Education under the research grant N N516 4307 33.

References

[1] D. Gambetta, “Can we trust trust”, in *Trust: Making and Breaking Cooperative Relations*. Oxford: Basil Blackwell, 1988, pp. 213–237.

[2] I. Pinyol and J. Sabater-Mir, “Arguing about reputation the lrep language”, in *Eighth Ann. Int. Worksh. Eng. Soc. Agents World ESAW’07*, Athens, Greece, 2007, pp. 192–207.

[3] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, “The eigentrust algorithm for reputation management in p2p networks”, in *Proceedings of the 12th International Conference on World Wide Web WWW’03*. New York: ACM, 2003, pp. 640–651.

[4] M. Schillo, M. Rovatsos, and P. Funk, “Using trust for detecting deceitful agents in artificial societites”, *Appl. Artif. Intel. J.*, vol. 14, no. 8, pp. 825–848, 2000.

[5] R. R. Yager, “On ordered weighted averaging aggregation operators in multicriteria decision making”, *IEEE Trans. Syst. Man Cybern.*, vol. 18, no. 1, pp. 183–190, 1988.

[6] R. Aringhieri, E. Damiani, S. De Capitani Di Vimercati, S. Paraboschi, and P. Samarati, “Fuzzy techniques for trust and reputation management in anonymous peer-to-peer systems”, *J. Amer. Soc. Inform. Sci. Technol.*, vol. 57, pp. 528–537, 2006.

[7] J. Sabater-Mir and M. Paolucci, “On representation and aggregation of social evaluations in computational trust and reputation models”, *Int. J. Appr. Reas.*, vol. 46, no. 3, pp. 458–483, 2007.

[8] V. Torra, “The weighted OWA operator”, *Int. J. Intel. Syst.*, vol. 12, pp. 153–166, 1997.

[9] W. Ogryczak and T. Śliwiński, “On optimization of the importance weighted OWA aggregation of multiple criteria”, in *International Conference Computational Science and Its Applications ICCSA 2007. LNCS*, vol. 4705. Heidelberg: Springer, 2007, pp. 804–817.

Michał Majdan received the M.Sc. degree in computer science from the Warsaw University of Technology, Poland, in 2003. Currently he prepares his Ph.D. thesis in computer science from the Institute of Control and Computation Engineering at the Warsaw University of Technology. He is employed by the National Institute of Telecommunications in Warsaw. He has taken part in projects related to data warehousing and analysis for a telecommunication operator. His research focuses on modeling, decision support, trust and reputation management.

e-mail: Michal.Majdan@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

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

Paper

An Eclectic Approach to Network Service Failure Detection Based on Multicriteria Analysis with an Example of Mixing Probabilistic Context Free Grammar Models

Paweł Białoń

Abstract—A method of failure detection in telecommunication networks is presented. This is a meta-method that correlates alarms raised by failure-detection modules based on various philosophies. The correlation takes into account two main characteristics of each module and the whole meta-method: the percentage of false alarms and the percentage of omitted failures. The trade-off between them is tackled with aspiration-based multicriteria analysis. The alarms are correlated using linear classification by support vector machines. An example of the profitability of correlating alarms in such way is shown. This is an example of probabilistic context free grammars (PCFGs), used to model the proper runtime paths of network services (and thus usable for detecting an improper behavior of the services). It is shown that the linearly mixing PCFGs can add context handling to the PCFG model, thus augmenting the capabilities of the model.

Keywords— failure detection, linear separation, probabilistic context free grammars, support vector machine.

1. Introduction

The domain of automatic failure detection in telecommunication and computer networks, becoming an extensively exploited domain, distinguishes with an exceptional variety of approaches used [1]. A broad spectrum of statistical/stochastic methods are used in failure detection, so are signal processing, discrete time sequence analysis, finite state machine methods, automatic reasoning, data mining, various classifiers, e.g., based on neural nets. The multitude of the existing approaches best proves that none of them is perfect. By choosing one of them, a designer of a network monitoring tool has to strongly narrow the area of a successful application of the tool.

This paper presents a concept of a meta-tool capable of integrating very different approaches known from the literature.

The proposition assumes an open architecture of the proposed tool – new literature approaches could be implemented as new modules of the tool. The indications of various modules are correlated, yielding a much more reliable assessment of the network state. Interesting is the way of correlating indications obtained from modules of completely different philosophies. The correlation uses linear classification and multicriteria analysis (we describe each module with two criteria that seem to be common throughout various approaches: the percentage of overlooked failures and the percent of false alarms). Several auxiliary hard optimization and simulation problems: large-scale, nondifferentiable, nonconvex are obtained. We propose to simplify some of them before solving, using statistical methods.

A fundamental question arises whether it is reasonable and useful to make linear combinations of outputs from various detection procedures. These procedures themselves may be described in languages strongly differing from "linear combining" – like some discrete approaches. To support our approach, we use a very interesting example. We mix indication from two modules detecting failures based on probabilistic context free grammar (PCFG) analysis of runtime paths [2]. By mixing them we essentially enlarge the expressiveness which a single module had: we add a context to the used grammar!

We have to make the area of application of our proposition more precise. It includes an automatic *failure detection*, where the management tool signalizies that a failure of the network is present and possibly gives some rough information of a type of the failure. The presented methodology can be applied to detecting both *service failures* and strict *network failures*. Though the paper more precisely analyzes some service failure approaches, we will refer jointly to both the types of failures using the short term of "network failures". Also, our tool would be suitable for a broader domain of *anomaly detection*, where an anomaly is understood in a broader sense that a failure (hardware, network-software or network-service) can express also untypical user behavior, connected with malicious activities, possible intrusions, frauds. Switching to making *proactive* failure of anomaly detections would be possible, by making some simple technical extensions, like shifting relevant time sequences within the tool, during its learning phase. However, there are bold challenges of *failure localization* (*reasoning about the failure reasons*) and *automatic* or *semi-automatic failure repairing* in which our tool would not acquit itself well. These are tasks by nature not well-suited for supervised learning methods, to whom our method ranks. These tasks are solved by expert systems and other artificial intelligence approaches.

In Section 2 we shall try to show the variety of existing failure detection approaches. The structure of our tool and the complex problem of tuning it will be discussed in Section 3. A discussion of the soundness of the described approach, together with the conclusions from our work is given in Section 4.

2. The Variety of Detection Approaches

We shall give a flavor of the variety of existing literature approaches to automatic network failure detection.

Various methods use various data about the network, coming from various sources. Let us, however, stress the increasing role of the simple network management protocol – SNMP (see [1] and the references therein) in acquiring such data. The network state is described by several tens of thousands variables (so called management information base variables – MIB variables), which can be probed at regular time bases. SNMP delivers variables connected with the traffic in particular network arcs, TCP/IP (transmission control protocol/Internet protocol) information (traffic, number of open connections, number of packets accompanying opening and closing connections, or accompanying some errors). Other useful sources of data are system logs, e.g., regarding line commands given by users (the logs are used in intrusion detection).

Let us enumerate some important classes of methods used in automatic failure detection.

1. Many approaches base on signal processing methods and stochastic methods. Usually, data about the network traffic in various moments are time sequences, often treated as realizations of stochastic processes.

   Such approaches can be often depicted as systems with elements like a filter, a statistic estimator, a discriminator, an alarm generator. The "systems" more or less accurately follow the structure from Fig. 1.

   An *alarm* is understood as a warning about incorrectness of the work of the network. An alarm is, however, easy to obtain and a rose alarm cannot itself prejudge that a failure is present. Alarm generation can be merely caused by exceeding some (lower or upper) limit value of the traffic intensity in some network arc. Similarly, the excess of some error frames rate may be examined and, in more advanced methods, the excess of some thresholds of certain signal statistics, calculated by an estimator.

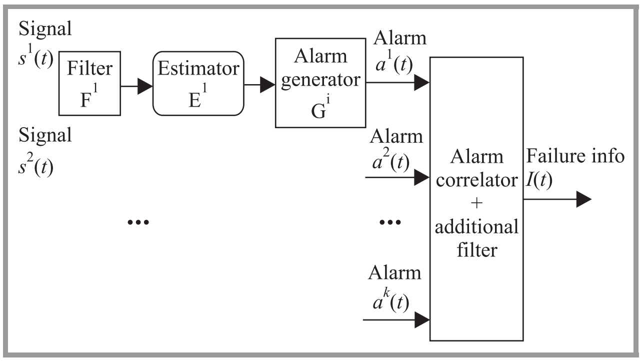


Fig. 1. A typical structure of the tools based on alarms generated from continuous traffic time sequences.

   A single alarm, when obtained in a simple way, is not very reliable in detecting failures (e.g., it may be false). Usually failures cause several alarms (e.g., simultaneously, a decrease of the traffic intensity and an increase of the error rate). The *alarm correlator* is an element obtaining the information on the presence of particular alarms and, based on it, deciding whether a failure is present (or localizing the failure – in systems that are capable of doing it). In particular, an occurrence of a single alarm at a time, can be ignored by the correlator.

   A pure value of some signal at some time may be not very relevant in raising or correlating alarms. For example, we would probably want to ignore some short-term incorrectnesses of the signal. Thus the described systems are often equipped with numerous *filters*, transforming signals both before and after the alarm generation.

   An example of the class of methods being discussed is presented in [3]. The traffic intensity in a certain network arc plays the role of signal $s^i(t)$ at Fig. 1. This signal is filtered (integrated within some time window) to reject temporary incorrectnesses. The obtained integral is some stochastic process; its distribution is modeled and estimated on a simple basis of ranking its values into several predefined intervals. When the signal goes out of a certain confidence interval (the intervals are different for various times of a day), an alarm is generated (if the situation is not only temporary – one more low-pass filter is applied).

   An interesting approach is presented in [4]. Each input signal s^i corresponds to a different MIB variable, usually representing traffic in a different layer: TCP, IP, data link. The idea is that anomalies propagate through the network layers, thus should be observable from different variables. Filter F is an autoregressive (AR) filter of rank 1. The estimator, in turn, calculates the defection of some simple statistics of the filter output (which are based on the variance) from the reference statistics (obtained from observations made in the immediate past – in some time window spreading up the present). A big defection means an "abrupt change" in the statistic properties of the input signal and thus – a probable occurrence of a failure. The volume of the defection corresponding to input s^i is expressed by a continuous alarm a^i from interval $[0,1]$. (Value 1 means the strongest defection). So we have continuous alarms, instead of zero-one. Correlating alarms is done by calculating

$$aBa^T\,,$$

where $a = \begin{bmatrix} a^1 \\ a^2 \\ \dots \\ a^k \end{bmatrix}$, B is a symmetric $k \times k$ matrix. When this value exceeds some threshold, a failure is ascertained. The presence of matrix A, with $\mathcal{O}(k^2)$ coefficients, already allows to model quite complex correlations between the alarms.

2. An example of failure detection by discrete methods is given in [5]. The authors have a method of discovering dependencies between two or more discrete time sequences, called multi-stream dependence detection (MSDD). This method can be applied to the problem of network failure proactive detection. The evolution of the network state (some input signal or some set of trivially obtained alarms) is described as a sequence of discrete values, corresponding to consecutive time moments, for example,

   ```
   1 5 7 3 6
   ```

   We can have two training sequences: one representing the presence or absence of a failure, another – some network signal. Then foreseeing a failure corresponds to finding correlation between elements of the sequences.

   Finding correlation in MSDD bases on templates, e.g., `*` in a template means "any value". Continuing our example, in the first sequence let us denote a presence of a failure by `F`, a normal network state – by `N`. Let us have positive integers in the second sequence, coding the values of some quantity measured in the network. MSDD can, for example, find a rule of the form:

   ```
   N N N N N N N N N F
   * 2 1 * * * * * * *
   ```

   The rule says that, the occurrence of the event consisting in an immediate (in one time instance) transition of the observed quantity from 2 to 1 indicates a presence of failure in 7 time units.

   Certainly, also more realistic, more complicated rules can be obtained. The algorithm searches for rules that have an outstanding support in the training data. The algorithm uses a mere Bayesian apparatus.

3. While the researchers prefering methods of signal analysis concentrate more on the alarm generation, the artificial-intelligence experts more willingly deal with the alarm correlation and also try to point out the reasons of the failure.

   In many works, like [6], *dependency graphs* are used to describe the propagation of a failure in the network. Some particular elements of the network are distinguished (a particular device, protocol, service, server, etc.). They are drawn as the graph nodes. If a malfunction of element A causes a malfunction of element B with probability p, we draw an arrow on the graph, from A to B, and p is denoted by the arrow. The propagation of a malfunction may be multi-stage, ending with an observed failure. Again, using the apparatus of conditional probabilities we can identify the most probable initial reason of the failure. In the cited paper, the way to do this leads through solving a combinatorial-optimization problem.

   Instead of using graphs, we can use logical expressions of some canonical form, of a lower nesting level [7]. Both the approaches need a laborious phase of obtaining the necessary knowledge from an expert. Both of them need a relatively hard updating of the monitoring software as the network changes in time (e.g., as it grows).

4. Detecting failures in remote databases (and other remote service environments) based on *run-time paths* is described in [2].

   A *runtime path* is composed of events happening in various places of the system. Events must have a common *request identifier* to be included into the same path. For example, an event can be a remote invocation of a procedure, data flow between remote components of the system, realization of a database query, spawning a thread by a Java application, etc. A request identifier can be a session identifier in the hypertext transfer protocol (HTTP).

   The simplest way to validate the correctness of the path is the analysis of the delays between events. The delays can have some reference distributions, build during an observation of a normal work of the distributed system. The conformance of delays currently being measured to these distribution can yield an assessment of the correctness of the system state.

   A much more powerful tool, so called probabilistic context free grammar can assess the correctness of the order of events on a path. PCFG (see [2] and the references therein) is an extension of mere context-free grammars, consisting in defining probabilities of the productions.

   A PCFG is a 5-tuple consisting of:

   - set of terminal symbols: $\mathfrak{T} = \{T^k : k = 1, \dots V\}$;
   - set of nonterminal symbols: $\mathfrak{N} = \{N^i : i = 1, \dots n\}$;
   - starting symbol $S \in \mathfrak{N}$;
   - set $\mathfrak{R} = \{R^j\}$ $(j = 1, \dots p)$ of productions of the form $N \rightarrow s$, where $N \in \mathfrak{N}$ and s is a finite sequence consisting of elements of $\mathfrak{T} \cup \mathfrak{N}$;
   - set of probabilities of productions $\mathfrak{P} = \{P^j\}$ $(j = 1, \dots p)$.

The probabilities of all productions with a given symbol on the left hand side must sum up to 1.

We can try to *derive* a given word (a sequence of terminal symbols) from the grammar, starting with the starting symbol and iteratively applying suitable productions until we end up with the word. For simplicity we assume that the derivation of any derivable word is unique.

Example 1: Let grammar G_1 be defined by $\mathfrak{N} = \{S, X, Y\}$, $\mathfrak{T} = \{a, b, c, d\}$, $\mathfrak{R} = \{$
$S \to XY \ (P = 1)$
$X \to a \ (P = 0.2)$
$X \to b \ (P = 0.8)$
$Y \to c \ (P = 0.5)$
$Z \to d \ (P = 0.5)$
$\}$

The derivation of word bc is following:
$S \to XY \to Xc \to bc$.
We used the 1st, the 3rd and the 2nd productions, in order.

The product of the probabilities of the used productions, 0.4 in the example, is the *probability of the word.* If symbols corresponded to events in our system (and words – to runtime paths) the probabilities of the word could be used to assess the correctness of a path (and of the distributed system state).

5. Certainly, other approaches are present. They may be based on "standard" methods (neural nets, other classification algorithms, clustering methods, the Markov process, etc.). Untypical approaches may prove usefulness. In [3], failures are suspected when the network devices are steering the traffic in a "strange" way, i.e., leading some arcs to an unnecessary saturation. In Fig. 2 the saturation threshold for each arc is 10 units, variables by arc denote the current traffic intensities. Then the configuration of intensities $x_{12} = 1$, $x_{23} = 1$, $x_{13} = 10$ is erroneous, arc $(1,3)$ is unnecessarily saturated, while a bypass exists through node 2. The reasoning seems simple on this simplistic example presented but becomes more sophisticated when we consider longer by-passes and distinguishes "commodities" in arcs, i.e., parts of the traffic with a particular sender and receiver.

   Another untypical approach uses the machinery of reference point multicriteria analysis [8].

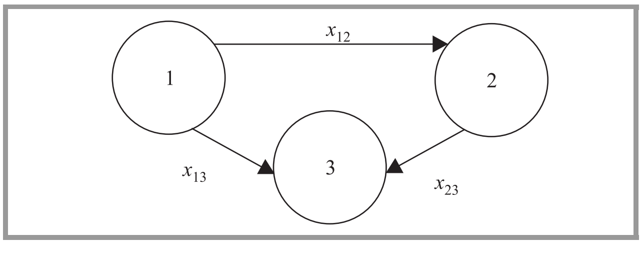


Fig. 2. An incorrectness of the traffic control.

3. The Idea of the Tool

3.1. Structure

Our hypothetical tool will contain failure detection modules of various philosophies. The coexistence of modules is possible due to their uniform treatment in the structure of the tool (Fig. 3).

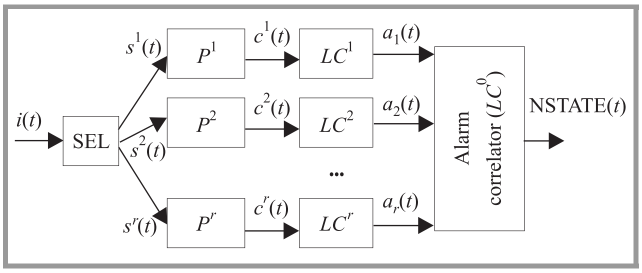


Fig. 3. Structure of the tool.

Each module consists of a *preprocessor*, into which the knowledge of a particular detection approach is coded, and a *linear classifier*. The input vector of the preprocessor (representing the current network state) is transformed to the output vector (preferably, a vector of reals), which describes the network state according to the knowledge of the approach. The output from the preprocessor, in the simplest case, would be the alarm value obtained due to the particular approach but will be rather some multidimensional description of the network state, and the role of the linear classifier would be to transform it into a continuous scalar representing the "normality level" of the network state.

Let us describe the elements of Fig. 3. We shall use a discrete time t – this variable will take natural values. Some of the quantities will be parametrized with this discrete time, thus they may be represented as functions of t. The main elements are:

1. Input signal vector $i(t) \in R^{n_i}$. Its coordinates may come, e.g., from MIB variables collected at time t; in general, they may represent diverse quantities.
2. Selector SEL – simple module selecting coordinates of i, from which vectors s^i are formed, used by particular preprocessors.
3. Preprocessors P^i for $i = 1, \dots r$, their respective outputs $c^i \in \mathbb{R}^{\eta_i}$, respective classifiers LC^0, ..., LC^r and alarms $a^i(t) \in \mathbb{R}$ rose by the classifiers for $i = 1, \dots r$. In general, each preprocessor remembers its inputs for at most H time instances, so

$$c^i(t) = c^i(s^i(t); s^i(t-1); \dots s^i(t-H+1)).$$

   There holds $a_i = \phi^i(c(t))$, with (operator ";" denotes the vector/number concatenation), $\phi : \mathbb{R}^{\eta^i}$ being an affine function:

$$\phi^i(x) = \omega^{i\top} x + \gamma^i, \qquad (1)$$

   where $\omega^i \in \mathbb{R}^{H\eta^i}$ and $\gamma^i \in \mathbb{R}$ are the classifier parameters, tuned in the learning phase, described later.

We treat the alarm value of -1 as de facto absence of an alarm, the value of 1 as rising an alarm.

4. Alarm correlator/LC^0. It receives at time t vector $a(t) = (a_1(t),\ a_2(t),\ \ldots,\ a_r(t))$. It returns a number greater than 0 (failure) or not greater than 0 (no failure). We do not equip the correlator with an additional low-pass filter eliminating "short-time failure indications" in our later reasoning, though it is desired in implementations.

3.2. The Case of Many Failures or Future Failures

The presented tool could be only simplistically extended to detect various types of failures, by multiplying the structure from Fig. 3 for separate failures. As already stated, we cannot expect great failure localization possibilities from our tool, which is of supervised learning tools class.

It is easier to augment our tool for the case of proactive failure detection. It suffices for modifying the learning phase, by shifting (in time) the teacher information of the state of the network (failure or no failure).

3.3. Teaching the Classifiers

Finding parameters ω^i γ^i of ith classifier is done in the supervised learning phase. We have historical examples $\mathfrak{e}_j$ $(j = 1,\ldots,n)$ of network states, where $\mathfrak{e}_j$ is an example from a historical time t^j, $e_j = c^i(t^j)$. For every example we know whether it is *positive*, i.e., describes a normal network state, or *negative*, i.e., describes a failure state.

Parameters ω^i γ^i, set to the solution of a certain optimization problem, are called support vector machine (SVM):

$$\underset{\omega^i \in \mathbb{R}^{\eta^i},\, \gamma^i \in \mathbb{R},\, y \in \mathbb{R}^n_+}{\text{minimize}} \quad \frac{1}{2}\|\omega\|^2 + Ce^\top y, \tag{2}$$

s. t.

$$-D\cdot(\mathfrak{e}^{j\top}\omega - \gamma) - y + 1 \le 0 \qquad \text{for } j = 1,\ldots n, \mathfrak{e}^j \text{ negative,}$$

$$(\mathfrak{e}^j\omega^\top - \gamma) - y + 1 \le 0 \qquad \text{for } j = 1,\ldots n, \mathfrak{e}^j \text{ positive.}$$

Here $e = (1,1,\ldots,1) \in \mathbb{R}^n$, $C, D > 0$. For the derivation of SVM problems see [9] or [10]. Learning parameter C controls the trade-off between the number of training examples misclassified by the taught classifier and the separation margin.[1] Learning parameter D is augmentation of the problem from [9] to control the relative impact of negative versus positive examples on γ^i and ω^i being found (it will mainly affect the later described trade-off between the tendency of the classifier to raise false alarms and to overlook failures). For a more thorough description of the above problem and a solver capable of solving it, see [11], [12].

3.4. Assessing the Modules Performance

Tuning the whole tool will be described in Subsection 3.5. For this, however, we must be able to assess the quality of tunnings (C, D) for a single module. This will be achieved by checking the attained compromise between two criteria: the number of false alarms and overlooked failures by the module. This is important because the whole tool will be assessed by similar criteria.

We divide some historical data about the network state into two sets: training data and test data. We teach a classifier with parameters C, D. We test the so taught module on the test data. We define function $q : \mathbb{R}^2 \to \mathbb{R}^2$ as

$$q(C,D) = \begin{bmatrix} \dfrac{\text{num. of misclassified positive test examples}}{\text{num. of positive test examples}} \\ \dfrac{\text{num. of misclassified negative test examples}}{\text{num. of negative test examples}} \end{bmatrix};$$

its coordinates are our *criteria* (the rate of false alarms, the rate of overlooked failures).

We make a parametric experiment – we teach our classifier for various combinations of C, D, i.e., for $(C,D) \in X$, where X is a finite subset of R^2_+. We obtain the following *attainable results set*:

$$Q = \{(y_1,y_2) \in \mathbb{R}^2 : \exists_{(c,d)\in X} q(c,d) = (y_1,y_2)\}.$$

We have to reject clearly unnecessary elements of Q, i.e., such elements that the classifier for some other setting of C, D gives one criterion not worse that in this element and the other criterion – better than in this element. By rejecting them we obtain the *efficient results set* (for some particular Pareto order):

$$Q^\star = \{(y_1,y_2) \in Q : \neg\exists_{(z_1,z_2)\in Q} \\ (z_1 < y_1 \wedge z_2 \le y_2) \vee (z_1 \le y_1) \wedge (z_2 < y_2)\}. \tag{3}$$

Since during the later tuning of the whole tool will see the modules only in terms of elements of Q (attained results for the module) we shall need to be able to return from an element of Q to (some) setting (C,D) that yielded it. For this reason, the tool must now memorize the relation between the settings and the attained results.

In the further analysis it will be easier to number elements of $Q^\star$ with one variable. Let us number the elements of $Q^\star$ with index ϑ, $\vartheta = 1,2,\ldots|Q^\star|$, according to the growing value of the first coordinate.

Such numbering is not ambiguous: there cannot exist two elements of $Q^\star$ with identical first coordinates and different second coordinates: definition (3) does not allow this.

Now each point in $Q^\star$ may be the value of a function $\varkappa$ of this index:

$$Q^\star = \{(y_1,y_2) = \varkappa(\vartheta) : \vartheta = 1,2,\ldots l\},$$

where $l = |Q^\star|$ and $\varkappa : \{1,2,\ldots l\} \to R^2$.

Remark 1: The coordinates of function $\varkappa$ are monotone: $\varkappa_1$ is an increasing function, $\varkappa_2$ is a decreasing function.

The increasing character of $\varkappa_1$ follows from numbering of the elements of Q by the first coordinate. The decreasing

[1] The existence of a (big) separation margin causes that the values of ϕ^i for the training examples are (much) isolated from 0. With a big separation margin we can hope for good generalization properties of our classifier.

character of $\varkappa_2$ follows from the numbering and from the constriction (3) of set $Q^\star$.
Finally, each module is characterized with its function $\varkappa$.

3.5. Teaching the Alarm Correlator

The alarm correlator yields $\mathtt{NSTATE}(t)$=$\omega^\top(a^1(t), a^2(t), \ldots a^r(t))$+$\gamma$, where $\omega \in \mathbb{R}^r$ and $\gamma \in \mathbb{R}$ are tunable parameters, ($\mathtt{NSTATE}(t) \leq 0$ means a failure, $\mathtt{NSTATE}(y) > 0$ – no failure).
We have been describing some quantities for a single module/classifier. Since we now have to consider all the modules jointly, we equip these quantities with an additional index i denoting the module number (i will run from 1 to r). So we shall make the following transformation in out notation:

$$l \to l_i,\ \vartheta \to \vartheta_i,\ \varkappa_j \to \varkappa_{i,j},\ C \to C_i,\ D \to D_i.$$

Moreover, C, D and ϑ will be vectors now: $C = (C_1, C_2, \ldots C_r)$, $D = (D_1, D_2, \ldots D_r)$, $\vartheta = (\vartheta_1, \vartheta_2, \ldots \vartheta_r)$.
We introduce function $\mathfrak{q}(t, \omega, \gamma)$ assessing the whole system:

$$\mathfrak{q}(C, D, \omega, \gamma) = \begin{bmatrix} \frac{\text{num. of positive test examples misclassified by correlator}}{\text{num. of positive test examples}} \\ \frac{\text{num. of negative test examples misclassified by correlator}}{\text{num. of negative test examples}} \end{bmatrix}.$$

The author would like to thank Prof. Wierzbicki and Dr. Granat for pointing out importance of assessing the system by such two criteria (defined by the coordinates of $\mathfrak{q}$). The choice of the compromise can be left to the network administrator. In the sequel we shall allow this choice with the apparatus of the reference point methodology [13].

We shall try to minimize both the criteria. To merit the "levels of achievement" in the minimizations we shall introduce a *scalarizing function* $s_{\bar{\bar{\mathfrak{q}}},\bar{\mathfrak{q}}} : \mathbb{R}^2 \mapsto \mathbb{R}$, [13] with parameters $\bar{\bar{\mathfrak{q}}} \in \mathbb{R}^2$ (the vector of so called reservation levels) $\bar{\mathfrak{q}} \in \mathbb{R}^2$ (the vector of so called aspiration levels). The reservation level for a criterion (a coordinate of $\mathfrak{q}$) is defined as such that the user does not want the criterion to deteriorate below the level. The aspiration level is defined as such that the user does not demand the criterion beyond the level.
The user can change $\bar{\bar{\mathfrak{q}}}$, $\bar{\mathfrak{q}}$ and the system solves the following optimization problem:

$$\underset{C, D, \omega, \gamma}{\text{maximize}}\, s_{\bar{\bar{\mathfrak{q}}},\bar{\mathfrak{q}}}(\mathfrak{q}(C, D, \omega, \gamma)), \tag{4}$$

finding the settings C, D, ω, γ.

3.5.1. The Case of a Few Modules

When there are only a few modules, problem (4) can be solved directly by a parametric experiment. Taking various combinations of the values of C, D, ω, γ, one can examine the resulting values of $\mathfrak{q}_1$ i $\mathfrak{q}_2$ by a direct simulation of the work of the modules and based on this one can calculate the values of the scalarizing function $s_{\bar{\bar{\mathfrak{q}}},\bar{\mathfrak{q}}}$, eventually choosing the combination of C, D, ω, γ that gave the biggest $s_{\bar{\bar{\mathfrak{q}}},\bar{\mathfrak{q}}}$.

3.5.2. The Case of Numerous Modules

The number of combinations of the values of C, D, ω, γ grows exponentially with the number of modules under a given sampling density. If there are more than several modules, the computations become unrealistic. Then, however, we can try to compute $\mathfrak{q}$ with statistical methods, using the central limit theorem.
For this we must assume that mistakes of particular modules are independent events. This assumption, a bit disputable, can be substantiated with the difference of the philosophies of the modules.
We shall consider two cases.

Some test example corresponds to a failure. We shall calculate the probability of misclassifying the example by the correlator (i.e., by the whole system).
We treat a_i as independent, discrete probabilistic variables, where $a_i = -1$ with probability $(1 - p''_i)$, and $a_i = 1$ with probability p''_i. We have denoted $p''_i = \varkappa_{i,2}(\vartheta_i)$. Recall that ϑ_i is a parameter indexing the set of efficient results for the ith module. Later ϑ_i will be made variables for each i – they will become decision variables in an optimization task that will serve for tuning the correlator.
We have

$$\mathrm{E}a_i = 2p''_i - 1 \quad \text{and} \quad \mathrm{Var}a_i = 4(p''_i - p''^2_i).$$

So for the probabilistic variable $\omega^\top a$ (i.e., for $\sum_i(\omega \cdot a_i)$) we have

$$\mathrm{E}(\omega^\top a) = \sum_i \omega_i(2p''_i - 1) \quad \text{and} \quad \mathrm{Var}(\omega^\top a) = 4\sum_i \omega_i^2(p''_i - p''^2_i).$$

We assume that the probabilistic variable $\omega^\top a$, being a sum of many independent variables has the distribution of[2]

$$N\left(\sum_i \omega_i(2p''_i - 1), \sqrt{\sum_i 4\omega_i(p''_i - p''^2_i)}\right),$$

where $N(\varepsilon, \sigma)$ denoted the normal distribution with expected value ε and variance σ^2.
The probability of overlooking the failure by the system is

$$\begin{aligned} P(\omega^\top a\gamma > 0) &= P(\omega^\top a > -\gamma) \\ &= 1 - D\left(\sum_i(\omega_i(2p''_i - 1)), \sqrt{\sum_i 4(\omega_i(p''_i - p''^2_i))}\right)(-\gamma) \\ &= 1 - D(0,1)\left(\frac{-\gamma - \sum_i(\omega_i(2p''_i - 1))}{4\sum_i(\omega_i^2(p''_i - p''^2_i))}\right), \end{aligned} \tag{5}$$

where $D(\varepsilon, \sigma)$ denotes the cumulative distribution function (CDF) of the normal distribution with expected value ε and variance σ^2.

Some test example corresponds to a correct network state. We put $p'_i = \varkappa_{i,1}(\vartheta_i)$. Using a similar reasoning as

[2] Since the variables have different variances, one should assure that none of the variances dominates the others, so that the conditions of the Linderberg-Feller theorem are satisfied. At least, the modules should be tuned similarly, in the sense that they yield misclassification rates of a similar rank.

above, we can calculate the probability of misclassifying this example by the system:

$$
\begin{aligned}
P(\omega^\top a\gamma \leq 0) &= P(\omega^\top a \leq -\gamma)\\
&= D\left(-\sum_i(\omega_i(2p'_i-1)), \sqrt{\sum_i 4(\omega_i(p'_i-p'^2_i))}\right)(-\gamma)\\
&= D(0,1)\left(\frac{-\gamma+\sum_i(\omega_i(2p'_i-1))}{4\sum_i(\omega_i^2(p'_i-p'^2_i))}\right). \qquad (6)
\end{aligned}
$$

Finally, in order to find the optimal tuning of ϑ, γ, ω we solve the following optimization problem:

$$\underset{\vartheta,\omega,\gamma}{\text{maximize}}\, s_{\bar{\bar{\mathfrak{q}}},\bar{\mathfrak{q}}}(\mathfrak{q}_1(t,\omega,\gamma),\mathfrak{q}_2(t,\omega,\gamma)),$$

where $\mathfrak{q}_1(\vartheta,\omega,\gamma)$ is given by (6) and $\mathfrak{q}_2(t,\omega,\gamma)$ – by (5). This problem is nondifferentiable due to the nondifferentiability of the scalarizing function $s_{\bar{\bar{\mathfrak{q}}},\bar{\mathfrak{q}}}$ and possible nondifferentiability of the necessary representation of $\varkappa_i$. It seems, however, optimistic that it has as quasi-analytical form, i.e., to compute the value of the goal function for a given argument one does not need to run preprocessors neither use the historical examples. Finding an effective technique of solving the problem is the subject of further research. It may be helpful that functions $\varkappa_i(\vartheta_i)_j$ and $D(0,1)(\cdot)$ are monotone and weights w_i can be assumed positive (if the modules are reasonable their votes should be taken with positive weights).

4. Discussion of the Proposition Soundness and Conclusions

Implementing and validating the proposition given in this paper is a large undertaking, involving an implementation of several of tens of preprocessors and also organizing the supervised learning, solving quite hard optimization problems, etc. Thus such undertaking is a subject of the further research and here we shall only give some partial arguments validating our approach.

The first, fundamental question is whether it is reasonable to combine particular methods from the literature (which can be very complex and subtle) with the quite rough tool of weighted summing. A very interesting example regarding the runtime path method with PCFGs, supports such combining. We shall show that by weighted summing of the "probabilities of the words" we can extend the expressiveness of the method using PCFGs by context handling!

That PCFG are context-free can be expressed as follows: the probability of a parsing subtree depends neither on earlier symbols nor on later symbols (this is a disadvantage of PCGGs, since many events in reality depend on the context [2]). Let us come back to Example 1 from Section 2. The probability of the one-node subtree $Y \rightarrow c$ equals to 0.5 independent of which the 1st symbol in the word is. So the occurrence of symbol c is independent of what has happened before (i.e., whether there was a or b in the first position) and amounts to 0.5. In other words, events "the 1st symbol in the word is a" and "the second symbol of the word is b" are independent, which can be written as follows:

$$P_{G_1}(ac) = P_{G_1}(a\star)\cdot P_{G_1}(\star c),$$

where $\star$ denotes any symbol allowed by the grammar at the given position.

Let us define grammar G_2, very similar to G_1 from Example 1 in Section 2 (even identical with G_1 in structure): $\mathfrak{N} = \{S, X, Y\}$, $\mathfrak{T} = \{a, b, c, d\}$, $\mathfrak{R} = \{$
$S \rightarrow XY$ $(P = 1)$
$X \rightarrow a$ $(P = 0)$
$X \rightarrow b$ $(P = 1)$
$Y \rightarrow c$ $(P = 0)$
$Z \rightarrow d$ $(P = 1)$
$\}$

Certainly, for G_2 there also holds the independence of the relevant events:

$$P_{G_2}(ac) = P_{G_2}(a\star)\cdot P_{G_2}(\star c).$$

However, under mixed probability, e.g., $P(.) \equiv 0.5P_{G_1} + 0.5P_{G_2}$, events "$a\star$" and "$\star c$" are no more independent. Namely, we have

$$P(a\star) = 0.5P_{G_1}(a\star) + 0.5P_{G_2}(a\star) = 0.5\cdot 0.2 + 0.5\cdot 0 = 0.1,$$

$$P(\star c) = 0.5P_{G_1}(\star c) + 0.5P_{G_2}(\star c) = 0.5\cdot 0.5 + 0.5\cdot 0 = 0.25,$$

$$P(ac) = 0.5P_{G_1}(ac) + 0.5P_{G_2}(ac) = 0.5\cdot 0.1 + 0.5\cdot 0 = 0.5$$

and finally:

$$P(ac) \neq P(a\star)\cdot P(\star c).$$

Mixing the probabilities introduced the desired contextual information handling to our tool.

Some preliminary experiments with teaching modules have been also performed (see [12] for details). The main outcome is an assessment of the form of set Q, giving flavor of what trade-offs between overlooking failures and raising false alarms can be obtained.

An experimental module had to detect failures consisting in breaking one of the arc of the skeleton computer network of the National Institute of Telecommunications. The module had very limited information about the current network state: as a single example, it had only a sequence of traffic intensities in some other arc at 20 consecutive time moments. To make the job of the module more difficult, the sequence was normalized so as its variance was drawn to 1 and its expected value was drawn to 0 for each example. So the module could only analyze the most subtle properties of the 20-element time sequences. It did it using a simple auto-regressive filter of rank 4 as the preprocessor.

The obtained set Q is shown in Fig. 4. For comparison, a line representing the behavior of the *random classifier* was built in the figure. The random classifier classifies each testing example as positive with probability p or as negative – with probability $1-p$ (independently of the real

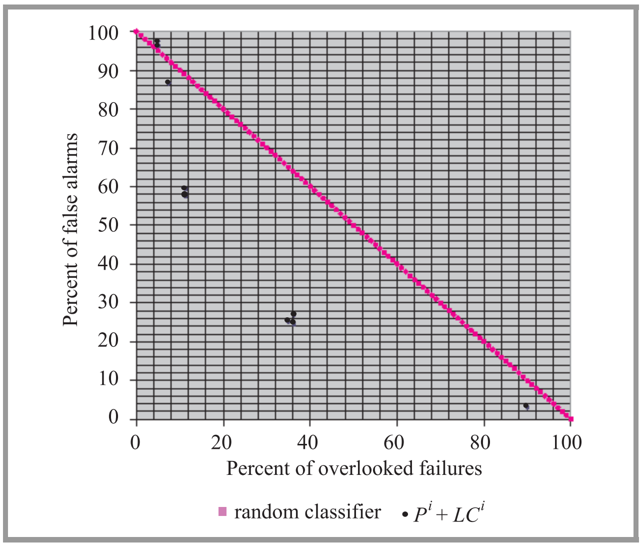


Fig. 4. Attainable results for the AR module.

current network state). By varying p we obtain the whole line in the graph.

Even under a difficult task posing, the AR module exhibited an efficiency clearly better than the random classifier. Correlating several tens of modules of a similar quality would probably be effective.

In conclusion, let us state that it is conceptually possible to join the efforts of various literature detection methods, of which no one is perfect. The main idea of the tool, combining even sophisticated detection methods known from the literature with the mere linear classification seems to be useful in some cases. The most important matters of the further research seem to be: solving the resulting optimization problems, incorporating some at least very rough classification of failures, making a thorough experimental validation.

References

[1] I. Katzela and M. Schwartz, “Schemes for fault identification in communication networks”, *IEEE/ACM Trans. Netw.*, vol. 3, pp. 753–764, 1995.

[2] M. Chen, A. Accardi, E. Kcman, J. Lloyd, D. Patterson, A. Fox, and E. Brewer, “Path-based failure and evolution management”, in *Proc. First Symp. Netw. Syst. Des. Implem. NSDI*, San Francisco, USA, 2004.

[3] J. Granat, W. Traczyk, C. Głowiński, J. Pietrzykowski, P. Białoń, and P. Celej, “Eksploracja i analiza danych pozyskiwanych z obiektów sieci teleinformatycznej dla wspomagania zarządzania i podejmowania decyzji”. Report 06.30.001.1, National Institute of Telecommunications, Warsaw, Dec. 2001 (in Polish).

[4] M. Thottan and J. Chuanyi, “Anomaly detection in IP networks”, *IEEE Trans. Sig. Proces.*, vol. 51, no. 8, pp. 2191–2204, 2003.

[5] T. Oates and P. R. Cohen, “Searching for structure in multiple streams of data”, in *Proc. Int. Conf. Mach. Learn.*, Bari, Italy, 1996, pp. 346–354.

[6] D. Gürer, I. Khan, and R. Ogier, “An artificial intelligence approach to network fault management”, SRI International, 1996.

[7] G. Jakobson and M. D. Weissman, “Alarm correlation – creating multiple network alarms improves telecommunications network surveillance and fault management”, *IEEE Network*, pp. 52–59, Nov. 1993.

[8] J. Granat and A. P. Wierzbicki, “Objective classification of empirical probability distributions and the issue of event detection”, in *Proc. 23rd IFIP TC 7 Conf. Syst. Modell. Optim.*, Cracow, Poland, 2007.

[9] D. R. Musicant, “Data mining via mathematical programing and machine learning”. Ph.D. thesis, University of Wisconsin – Madison, 2000.

[10] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[11] P. Białoń, “A linear support vector machine solver for a huge number of training examples”, *Contr. Cybern.* (to appear).

[12] P. Białoń, A. P. Wierzbicki, and J. Granat, “Narzędzia monitoringu w zarządzaniu sieciami komputerowymi i bazami danych”. Report 06.30.003.7, National Institute of Telecommunications, Warsaw, Dec. 2007 (in Polish).

[13] J. Granat and A. P. Wierzbicki, “Multicriteria analysis in telecommunication”, in *Proc. 37th Ann. Hawaii Int. Conf. Syst. Sci. HICSS’04*, Hawaii, USA, 2004, track 3, vol. 3.

Paweł M. Białoń was born in Warsaw, Poland, in 1971. He received his M.Sc. in computer science from the Warsaw University of Technology in 1995. He is with the National Institute of Telecommunications in Warsaw. His research focusses on nonlinear optimization methods and decision support, in particular on projection methods in optimization and support vector machine problems. He has participated in several projects applying decision support in various areas: telecommunications (techno-economic analyses in developing telecommunication networks, network monitoring), also in agricultural and environmental areas.

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

Paper

A Level-Based Approach to Prioritize Telecommunications R&D

Joana Fialho, Pedro Godinho, João Paulo Costa, Ricardo Afonso, and José Gonçalo Regalado

Abstract—In this paper, an approach to evaluate R&D projects in telecommunications is presented. These projects have particular features that cannot be properly incorporated by classical evaluation methods. This approach incorporates different criteria, both quantitative and qualitative, and also management flexibility and uncertainty. Thus, it is an approach that can be applied to real data of R&D projects in a telecommunications company.

Keywords— multicriteria decision support, R&D project evaluation.

1. Introduction

Research and Development (R&D) projects in telecommunications have particular features that cannot be properly incorporated by classical evaluation methods. A correct evaluation of these projects must consider different criteria, both quantitative and qualitative, and moreover, management flexibility and uncertainty. The purpose of this paper is to present an approach to evaluate R&D projects in telecommunications. In these projects, there is much uncertainty, especially associated with management flexibility: it is necessary to make decisions under an environment of uncertainty. For example, it is necessary to decide when to start a project or when to launch a product. In other occasions, it is important to decide if there are reasons to abandon a project. There are several methods that evaluate R&D projects; in this paper, we present an approach that takes into account the specific structure of a telecommunications company.

The approach presented in this paper considers two levels of decision: the activity and the aggregate. An activity is a set of tasks with specific objectives and characteristics. It is planned at short term. An aggregate is a set of connected activities that are oriented towards a specific product or service. It is planned in the medium term (normally, 4 or 5 years).

In R&D projects in telecommunications, there are different decision levels: top management level; aggregate level (project managers); or activity level (task managers). To incorporate and coordinate these decision levels, two structures were considered: one for activities and another for aggregates. These structures allow, at each level, the identification of the relevant criteria and their relative importance by the decision makers (DMs) of the respective level in the structure. This approach is inspired on the evaluation process developed for the British Aerospace Military Aircraft and Aerostructures, presented in [1]. To infer the relative importance of criteria in each level, DMs make comparisons among criteria. However, in the present approach, the number of comparisons is limited in order to avoid overloading the DMs with excessive information requests. This process is based on Harker [2] and on Saaty [3]. The evaluation of aggregates can be made one by one. If just one aggregate is under evaluation, it is compared against benchmark aggregates, previously defined by the top management. If several aggregates are evaluated individually against benchmark aggregates, the results can give a proxy of the attractiveness of each aggregate. If several aggregates are compared among themselves, a relative performance proxy can be obtained. These proxy values are the basis for allocating human resources among the aggregates, providing managers with a global guide when deciding which projects to pursue and the level of activity in each of those projects.

The activities structure also has different levels of criteria. For each level, the respective DMs identify the relevant criteria and its relative importance. Activities are also evaluated in a process based on comparisons among themselves. These comparisons are made in each criterion. A global performance index is obtained for each activity. Note that each activity belongs to an aggregate. This value is used to help to determine the human resources allocation level of each activity, taking into account the resources allocated to the respective aggregate.

To test this approach, a prototype was created with two files (one to evaluate aggregates and another to evaluate activities) in "Microsoft Excel" and some aggregates and activities were evaluated. The results were considered to reflect the company policy, which was captured through the information requests along the evaluation process.

This paper is organized as follows: in Section 2, an overview in evaluation of R&D projects is presented. The developed approach is presented in Section 3 and Section 4 concludes.

2. An Overview in Evaluation of R&D Projects

The approach developed in this paper was inspired on the evaluation process developed for the British Aerospace Military Aircraft and Aerostructures, presented in [1]. The need of a process based on technology management was recognized by Gregory [4]. This process takes into account specific areas of technology management, such as competence analysis, innovation R&D management, among others [1]. There are different methods to evaluate R&D projects, but it is difficult to aggregate all issues that characterize these kind of projects. Some of these methods are described in [5].

Economic models cannot take into account qualitative factors and treat each project in isolation [1]. Moreover, they also require solid financial data. Besides, traditional methods cannot incorporate management flexibility or some uncertainty factors. On R&D projects, new information may arrive and some changes on market conditions may take place. These aspects may lead to a change of strategy [6]. Real option methods incorporate both uncertainty and management flexibility [7], but it is difficult to apply them to real data, because of the complexity of the inference of some parameters.

There are other evaluation methods, such as the scoring method that evaluates each project in isolation or the comparative method that compares each project with another one or with a set of other projects. In the comparative method, different people can provide different comparisons, and evaluation may change over time [8].

Project evaluation can be made by mathematical programming, but it is difficult to incorporate uncertainty factors. Other difficulty concerns on the aggregation of different measures into a single value [9]. Charnes *et al.* [10] developed data envelopment analysis (DEA), a method based on linear programming that can incorporate variables with different units. Nevertheless, this method does not emphasize economical aspects neither uncertainty factors.

Multicriteria analysis can also be a tool to evaluate projects, where the projects are the alternatives which are being analyzed. The developed approach was inspired on the evaluation process developed for the British Aerospace Military Aircraft and Aerostructures, presented in [1]. Moreover, both quantitative and qualitative criteria were used. Boucher and MacStravic [11] also used quantitative criteria in a structure they constructed to evaluate R&D projects.

The structure used in this paper to evaluate R&D projects in telecommunications is based on the analytical hierarchy process (AHP) [12]. This methodology is used in many areas, including the evaluation of projects R&D. For example, Shin *et al.* [13] used this methodology to evaluate the national nuclear R&D projects in the case of Korea. Other example is given by Poh *et al.* [8]. They used AHP to compare methods that evaluate R&D projects.

Nevertheless, in the presented approach, some aspects of AHP were modified, in order to cope with some problems, like the number of comparisons and the integration of quantitative criteria. For a good review of AHP problems, see [14] and [15].

3. The Approach

The approach here presented results from meetings with telecommunications company staff. These meetings allowed to identify the information already available in the company, and the information that can be, reasonably, expected to be provided by project managers. The results of these meetings allowed to define the model structure. After constructing this structure, other meetings were made to define criteria and parameters that should be in evaluation model. Finally, through other meetings, criteria weights were defined.

Two levels of decision were identified to build the model and perform the evaluation of projects: the *activity* and the *aggregate*. An activity is a set of tasks with specific objectives and characteristics. It is planned at short term. An aggregate is a set of activities which are connected and guided to a specific product or service. The aggregate is planned in the medium term (normally, 4 or 5 years).

In the R&D sector of a telecommunications company, there are different decision levels: there are decisions made by the top management; project managers have to make decisions about the aggregates (aggregate decision level); task managers have to decide about activities (activity decision level). The proposed approach considers two structures to incorporate and coordinate these decision levels: one for the evaluation of activities and another for aggregates evaluation. Relevant criteria and their relative importance are identified at each level of both structures. These values must reflect institutional preferences. Thus, the criteria and their weights must be defined by the DMs of the respective level of the structure. The weights of the criteria reflect their importance in the category they belong to.

To infer the weights of criteria in each level, DMs make comparisons among criteria. However, the number of comparisons is limited to avoid requiring excessive information from the DMs. Thus, the DMs fill out part of a matrix of comparisons, $A = [a_{i,j}]_{i,j=1,\ldots,n}$, where n is the number of the criteria in a level of the structure, and a_{ij} represents the comparison between the criterion i and the criterion j. The DMs can fill out the matrix with the values that are defined in the following scale:

$$1/9,\ 1/7,\ 1/5,\ 1/3,\ 1,\ 3,\ 5,\ 7,\ 9,$$

where $a_{ij} = 1/9$, means that the criterion i is extremely less important than the criterion j; $a_{ij} = 1$, means that both criterion have the same importance; $a_{ij} = 9$, means that the criterion i is extremely more important than the criterion j.

The DMs can also use other numerical values that are not directly defined in the scale.

Suppose that $n = 4$ and the DMs filled out the matrix with the following values:

$$A = \begin{bmatrix} 1 & 2 & 1/3 & 3 \\ & 1 & & 4 \\ & & 1 & \\ & & & 1 \end{bmatrix}.$$

Note that $a_{ji} = 1/a_{ij}$, which allows the DMs fulfill just part of the superior (or inferior) triangle of the matrix. Besides, $a_{ii} = 1$, because it represents the comparison of one criterion with itself.

Through this matrix, it is possible to get weights of criteria, $w_i,\ i = 1,\ldots,n$ by minimizing the inconsistency index of the AHP method [2], [3]. In the case of the example, the weights obtained were $w_1 = 0.212,\ w_2 = 0.15,\ w_3 = 0.586,\ w_4 = 0.052$.

It is also possible to provide the complete matrix of comparisons and the respective inconsistency index. The rest of the matrix of comparisons is given by $a_{ij} = w_i/w_j$, in order to make the matrix consistent with the judgments already provided. The weights of the criteria and the complete matrix of comparisons are shown to the DMs, which allow them to maintain or revise their judgements.

In the case of the example, the complete matrix is

$$A = \begin{bmatrix} 1 & 2 & 1/3 & 3 \\ 0.5 & 1 & 0.26 & 4 \\ 3 & 3.89 & 1 & 11.29 \\ 1/3 & 0.25 & 0.09 & 1 \end{bmatrix}$$

and the inconsistency index is 0.031. If this index was too high (larger than 0.1), the DMs should revise their judgements.

Once the criteria and their weights have been defined, it is possible to evaluate both aggregates and activities in the different criteria. These evaluations are made by DMs and taking into account data provided by project managers. With these evaluations, it is possible to aggregate them into a single value that represents the global evaluation of an aggregate or an activity.

In a global manner, aggregates evaluation is used to allocate human resources. Activities evaluation is used to allocate human resources, inside the corresponding aggregate.

3.1. The Evaluation of Aggregates

The level structure to evaluate aggregates begins by specifying the type of aggregate, because the weights of criteria may be different for different kinds of aggregates. After some interviews, it was concluded that there were two types of aggregates: strategic ones with long term objectives; and business ones, aiming to obtain profits at a shorter term.

The second level of the structure includes criteria of a superior level or categories of objectives. After some meetings and respective analysis, three categories were considered: strategic, operational and financial. These categories include different criteria that were identified by top management.

The criteria identified in strategic category included the contribution to the company's image, strategic partnerships, market leadership, acquired skills, importance of company credibility for the client and importance of technology.

In the operational category, the criteria that were identified are technical, like technological uncertainty, scarcity of needed resources, solution flexibility, dependence on external entities and client satisfaction.

In the financial category, an indicator that reflects the value of the aggregate in a perspective of 4 or 5 years was used. In addition, it was recognized that other factors were important in this category, like expected loss for abandonment and postponement possibility. These factors are modeled as qualitative criteria. Thus, in this category, there are two qualitative criteria and a quantitative one. Note that the quantitative criterion can be well represented by net present value (NPV), because this measure reflects all cash flows predicted for the following 4 or 5 years.

However, there are aggregates where it is not possible to make forecasts at a such term, due to uncertainty factors. In this case, the quantitative criterion (NPV) is replaced by qualitative factors and by one quantitative factor that reflects the aggregate's value at a short term (1 year). This quantitative factor can be NPV, but with reference to the following year. So, in the case that is not possible to estimate NPV for 4 or 5 years, this criterion is replaced by a financial value for short term (NPV for 1 year), plus the qualitative criteria growth perspectives and market trend.

Figure 1 represents the structure of evaluation. Aggregates can be evaluated either in isolation or by comparing them with each other. The evaluation of one aggregate is made by comparing it with benchmark aggregates. These benchmarks are previously defined by the company administration. The financial values are defined in each benchmark aggregate both at short term and at long term. For the qualitative criteria, for each benchmark aggregate, the percentile relatively to real aggregates in the company is given. These benchmark aggregates are defined in order to make it possible to compare them with the aggregates under evaluation.

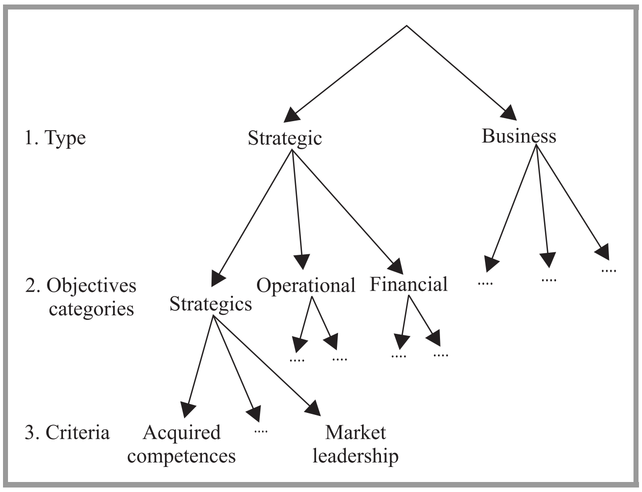


Fig. 1. Structure to evaluate aggregates.

For each criterion, the DMs fill out part of a matrix of comparisons, $A = [a_{ij}]_{i,j=1,\ldots,m}$, where m is the number of aggregates that are being evaluated, with values that compare the aggregate under evaluation against the benchmark aggregates. These values are defined in a scale that is constituted by the values

$$1/9,\ 1/7,\ 1/5,\ 1/3,\ 1,\ 3,\ 5,\ 7,\ 9,$$

where if $a_{ij} = 1/9$, then aggregate i is extremely worse than aggregate j in the criterion; $a_{ij} = 1$, then both aggregates have the same performance in the criterion; $a_{ij} = 9$, then aggregate i is extremely better than aggregate j in the criterion.

However, DMs can fill out each cell of the matrix with other values that are not in the scale. The process described previously for the inference of weights of criteria is also used in this case (based on [2], [3]). In this case, the process is used to infer the relative importance of each aggregate in the criterion. For each qualitative criteria, the relative importance of the aggregate under evaluation and of the benchmark aggregates are calculated. These values of relative importance provide a relative evaluation of aggregates in the criterion, i.e., a higher relative importance means a better performance in the criterion. With the process described on [2], [3], the inconsistency index and the complete matrix are calculated and shown to the DMs. This information allows DMs to realize if the comparisons they have introduced are coherent. If the index of inconsistency is very high or the complete matrix is, somehow, unexpected, the DMs can review their comparisons.

To evaluate the aggregate on the quantitative criterion, the aggregates (the aggregate under evaluation and the benchmark aggregates) are compared, through their financial values. These comparisons are based on weights of negative and null financial values defined before.

During the meetings, it was defined that negative and null financial values are not linear. This happens because, for example, a loss of 100 (NPV $= -100$) may not be half as bad as a loss of 200 (NPV $= -200$). However, it can be assumed that there is linearity for positive financial values.

Thus, company's representatives compare negative and null financial values, fulfilling a matrix of comparisons. With this matrix of comparisons (that is possibly incomplete) and applying the previous methodology for incomplete matrices of comparisons, the weights or impacts of those negative and null values are calculated.

With the aggregates financial values, a global matrix of comparisons is constructed. The purpose of this matrix is to compare the values of reference defined before and the financial values that represent the aggregates. Its construction is based on impacts of negative and null values defined previously, and on the assumption that there is linearity among positive values.

Thus, this global matrix includes the comparisons among negative and null values and the linear comparisons among the positive values.

With this global matrix, it is possible to calculate the impacts for all values that are being compared. Taking only values of aggregates (aggregate under evaluation and benchmark aggregates), it is possible to normalize them to get impacts or relative importance of aggregates in the quantitative criterion.

With the relative importance of the aggregate under evaluation in all criteria, it is possible to determine a value for the aggregate, taking into account the relative importance of the criteria and the importance of the categories of objectives.

If several aggregates are evaluated individually against benchmark aggregates, the results of each evaluation can give a proxy of attractiveness of each aggregate.

To evaluate a set of aggregates, the same methodology described before is used, i.e., the aggregates are compared among themselves in all criteria. After having the relative importance of the aggregates in all criteria, a global value is calculated for each aggregate (the value that includes the relative importance of the aggregates in all criteria weighed by the relative importance of criteria and the weights of the categories of objectives). So, the evaluation of a set of aggregates provides a relative performance proxy for each aggregate. These proxy values will be the basis for human resources allocation among aggregates, providing managers with a global orientation about decisions of human resources allocation. Moreover, it is possible, with these values, to have a global guide about the level of activity in those aggregates and an orientation if it is necessary to decide which projects to pursue.

3.2. Activity Evaluation

Activity evaluation is based on a similar approach to the one developed to evaluate aggregates. However, the structure has different levels of criteria. There are activities very different and, thus, different types of activities were identified:

- **Exploratory research**: these activities aim to explore and study new technologies to know if it is possible to achieve interesting results.
- **Experimental development**: these activities have a defined objective on a defined application. So, the investigation is guided to advance on a specific orientation.
- P**roduct development and engineer services**: the purpose of these activities is to develop a product for immediate sale, or to provide support to an existing product.

The type of activity (see Fig. 2) corresponds to the first level of the structure. The category of objectives is in the second level, similarly defined to the one used with the aggregates: strategic, operational and financial. These categories and respective criteria were identified through meetings with members of the company's administration. The identified criteria in the strategic category were: contribution to the company image, market leadership, acquired skills, strategic partnerships, company credibility for the client and importance of technology. In the operational category, the criteria that were identified were: technological uncertainty, scarcity of needed resources, importance of the activity to the respective aggregate and dependence from external entities. In the financial category, the long term perspective is not considered. In this way, the criteria presented and identified in this category were financial value at a short term, growth perspectives, market trend, expected loss for abandonment and postponement possibility.

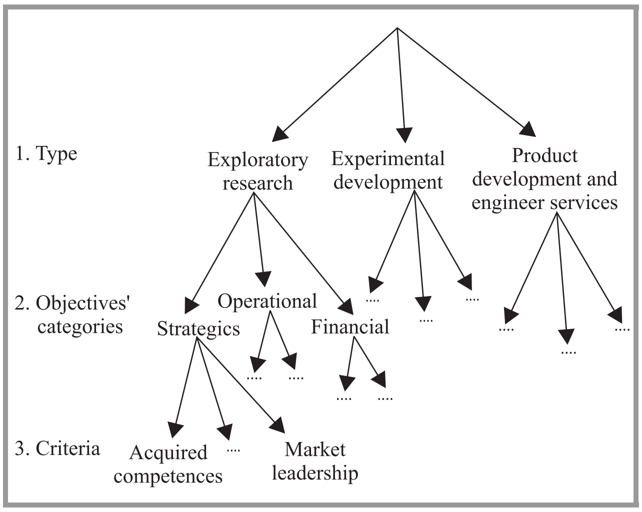


Fig. 2. Structure to evaluate activities.

The activities that belong to the same aggregate are compared among themselves. Through a process similar to the one used to evaluate a set of aggregates, a proxy value is calculated for each activity.

Note that each activity belongs to an aggregate and its evaluation is used to allocate human resources inside the aggregate. With the criteria weights and the performance of each activity on each criteria, it is possible to determine a global performance indicator or the evaluation of each activity. This indicator is the basis of human resource allocation decisions within the aggregate. For these decisions, one of the following procedures can be used.

- Defining global priorities for each activity, and selecting activities according to their priority. Restrictions may be used, like minimum and maximum values for human resources, for each type of activity.
- Mathematical programming to maximize the sum of priorities, taking into account some constraints, like limits on the number of human resources in each type of activity.

The activity and aggregate evaluation can give managers indications for resources allocation and allow them to gain sensibility for future decisions in the area.

3.3. The Prototype

A prototype was created, in "Microsoft Excel", and tested with data of a telecommunications company. Some aggregates and activities were evaluated, after getting the organizational preferences.

This prototype has different functionalities relatively to criteria. For the criteria, DMs can introduce new criteria or categories of objectives. It is possible to treat quantitative criteria (which are, mostly, financial) and the prototype guarantees their compatibility with the qualitative criteria.

The prototype computes the weights of criteria, through comparisons among them.

To evaluate aggregates or activities, the prototype allows to compare them with each other, in order to get their global evaluation. In addition, it is possible to introduce new aggregates or activities in the evaluation. If DMs want to evaluate one aggregate in isolation, they may use benchmark aggregates, which are defined in the prototype.

The evaluation provides a relative priority index for each aggregate or activity under evaluation. If one aggregate is being evaluated in isolation, a global priority index is calculated.

These indexes provided by the prototype allow DMs to support or justify the resources allocation of the aggregates and of their activities.

Some aggregates and activities were evaluated, taking into account the preferences of the company. The results of evaluation reflected these preferences and the company policy.

4. Conclusions

In this paper, it was presented an approach to evaluate telecommunications projects, taking into account two distinct levels: the activity and the aggregate. In each level, there are different types of decision, allowing the construction of a structure. These structures (one for the activities, another for the aggregates) allow the incorporation and coordination of the different decision levels presented in the activities and aggregates evaluation. In each level of the structure, different criteria are used. The decision makers responsible for each level define its criteria and respective weight.

The evaluation of aggregates can be made individually, through comparisons between the aggregate under evaluation and benchmark aggregates. If several aggregates are evaluated individually, the results can give a proxy of attractiveness of each aggregate. The evaluation of a set of aggregates may be made by comparing themselves with each other. The evaluation gives a global performance indicator for each aggregate.

The evaluation of activities is made through comparisons among a set of activities which belongs to a specific aggregate. A global priority index is given for each activity under evaluation.

A prototype was created to evaluate both aggregates and activities. These evaluations were based on data from a telecommunications company. In general, the results reflected the company policy. This policy was integrated in the evaluation process through the requested information.

The evaluation of aggregates and activities supports managers decisions on resources allocation. The presented approach may detect incoherences in evaluation when DMs have to compare criteria, aggregates or activities. On the other hand, this model may integrate new decisions when new opportunities arise that were not foreseen when the projects began. This integration also gives an orientation for resource reallocation. Finally, it is possible to identify the sources of evaluation errors, when such errors are detected.

With this approach, it is also possible to provide incentives to the identification of strategic opportunities and operational flexibility, through the definition of multiple criteria.

To conclude, the tool here presented may help to achieve better resource allocation decisions in a telecommunications company.

References

[1] C. Farrukh, R. Phaal, D. Probert, M. Gregory, and J. Wright, "Developing a process for the relative valuation of R&D programmes", *R&D Manage.*, vol. 30, no. 1, pp. 43–53, 2000.

[2] P. T. Harker, "Incomplete pairwise comparisons in the analytic hierarchy process", *Math. Model.*, vol. 9, no. 11, pp. 837–848, 1987.

[3] T. L. Saaty, *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*, vol. VI. Pittsburg: RWS Publ., 1994.

[4] M. J. Gregory, "Technology management – a process approach", *Proc. Instit. Mech. Eng.*, vol. 209, pp. 347–356, 1995.

[5] A. Henriksen and A. Traynor, "A practical R&D project-selection scoring tool", *IEEE Trans. Eng. Manage.*, vol. 46, no. 2, pp. 158–170, 1999.

[6] A. Dixit and R. Pindyck, *Investment under Uncertainty*. New Jersey: Princeton University Press, 1994.

[7] P. Godinho, J. Regalado, and R. Afonso, "A model for the application of real options analysis to R&D projects in the telecommunications sector", *Glob. Bus. Econom. Anth.* II, pp. 414–422, Dec. 2007.

[8] K. L. Poh, B. W. Ang, and F. Bai, "A comparative analysis of R&D project evaluation methods", *R&D Manage.*, vol. 31, no. 1, pp. 63–75, 2001.

[9] S. Coldrick, P. Longhurst, P. Ivey, and J. Hannis, "An R&D options selection model for investment decisions", *Technovation*, vol. 25, pp. 185–193, 2005.

[10] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision makig units", *Eur. J. Oper. Res.*, vol. 2, no. 6, pp. 429–444, 1978.

[11] T. O. Boucher and E. L. MacStravic, "Multiattribute evaluation within a present value framework and its relation to the analytic hierarchy process", *Eng. Econom.*, vol. 37, no. 1, pp. 1–32, 1991.

[12] T. L. Saaty, *The Analytic Hierarchy Process: Planning, Priority, Setting Resource Allocation*. New York: McGraw-Hill, 1980.

[13] C.-O. Shin, S.-H Yoo, and S.-J. Kwak, "Applying the analytical hierarchical process to evaluation of the national nuclear R&D projects: the case of Korea", *Progr. Nucl. Ener.*, vol. 49, pp. 375–384, 2007.

[14] A. Wierzbicki, J. Granat, and M. Makowski, "Discrete decision problems with large number of criteria", Interim Reports, IR-07-025, International Institute for Applied Systems Analysis (IIASA), 2007.

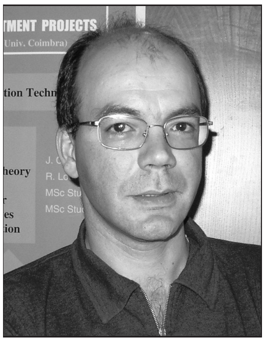
[15] C. Bana, J. Costa, and J. Vansnick, "MACBETH – an interactive path towards the construction of cardial value functions", *Int. Trans. Opl. Res.*, vol. 1, no. 4, pp. 489–500, 1994.

Joana Fialho is a Ph.D. student in management/decision aiding science, School of Economics at the University of Coimbra, Portugal. She is a lecturer in Polytechnic Institute of Viseu, Portugal, and a researcher at INESC – Coimbra. She graduated in mathematics from the Department of Mathematics, School of Science and Technology at the University of Coimbra, and a M.Sc. in information management, from School of Economics at the University of Coimbra. Her area of scientific activity is decision aid, namely evaluation of R&D projects in telecommunications.

e-mail: jfialho@mat.estv.ipv.pt
Institute of Computer and Systems Engineering (INESC)
Department of Mathematics
School of Science and Technology
University of Coimbra
Campus Politécnico de Repeses
3504-510 Viseu, Portugal

Pedro Godinho is an Auxiliary Professor with tenure at the Faculty of Economics, University of Coimbra, Portugal. He holds a Ph.D. in management and an M.Sc. in economics from the University of Coimbra. His research interests include Project Analysis and Evaluation, Real Options, Capital Markets Behaviour and Project Management. He is also a researcher at GEMF – Coimbra. He is the author or a co-author of several refereed publications in the areas of finance and operations research. He was a researcher of several funded research projects.

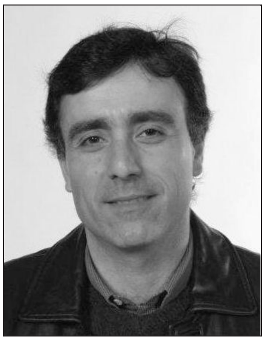
e-mail: pgodinho@fe.uc.pt
Monetary and Financial Studies Group (GEMF)
Faculty of Economics
University of Coimbra
Av. Dias da Silva no. 165
3004-512 Coimbra, Portugal



João Paulo Costa is a Full Professor at the Faculty of Economics, University of Coimbra, Portugal. His research interests include decision support systems, information systems, analysis and evaluation of projects and multicriteria decision analysis/making. He holds his Ph.D. in business economics from the University of Coimbra.

He is also a researcher at INESC – Coimbra. He is the author or a co-author of more than 100 refereed publications. He was the main researcher of various funded research projects.

e-mail: jpaulo@fe.uc.pt
Institute of Computer and Systems Engineering (INESC)
Faculty of Economics
University of Coimbra
Av. Dias da Silva no. 165
3004-512 Coimbra, Portugal



Ricardo Afonso received his B.Sc. degree in statistics and operations research from the Faculty of Sciences of the University of Lisbon, Portugal, in 1988. In 1998, he received the M.Sc. degree in information management in the organizations from the Faculty of Economy of the University of Coimbra. He joined Portugal Telecom Inovação (formerly CET) in 1988 for the development of telephone and multi-service telecommunications planning tools as well as the software development for the digital switching developed by the company. He was also active in this area, in the scope of RACE, RACE 2 and EURESCOM projects. More recently he has involved in coordination of IST projects (GMF4iTV and porTiVity). Currently he is head of the Innovation System and Projects unit and is responsible for supporting the planning and control activity in PT Inovação in addition with other specific project management activities.

e-mail: ptinovacao@ptinovacao.pt
PT Inovação
Rua Eng. José Ferreira Pinto Basto
3810-106 Aveiro, Portugal

José Gonçalo Regalado is a master student from Erasmus University – Rotterdam School of Management, where he is enrolled in the M.Sc. in business administration (finance and investments) program. He received his B.Sc. in management (4 years) from University of Coimbra (Faculty of Economics), Portugal, in 2007. Afterwards, he pursued an Executive Education Program in Entrepreneurship and Innovation Management from the Catholic University of Portugal (School of Economics and Management), where he received his executive education diploma in July of 2008. He was a member of the Talent Program of Portugal Telecom Inovação, where he was a part of the Innovation System and Projects Evaluation Division in the Planning, Control and Resources Department.

e-mail: ptinovacao@ptinovacao.pt
PT Inovação
Rua Eng. José Ferreira Pinto Basto
3810-106 Aveiro, Portugal

Paper

Ontology Creation Process in Knowledge Management Support System for a Research Institute

Cezary Chudzian

Abstract—Though the issue of knowledge management is a hot subject of interest in nowadays market companies, integrated solutions fit to the specific needs of research institutes still require more attention. This paper documents a part of the research activities performed at National Institute of Telecommunications, related to development of research institute knowledge management support system. The ideas lying in the background of the system come from the recent theories of knowledge creation and creativity support and from experience with everyday practice of knowledge management in market companies. Main focus is put here on the issue of creation of a research topics ontology that is meant to be a semantic backbone of the system. Three-stage approach is proposed, aiming at the construction of ontologies for different levels of organizational hierarchy, from individual researcher, through group or unit, up to the whole institute. Created ontologies are linked to knowledge resources and support diverse activities performed at those levels.

Keywords— creativity support, knowledge creation, ontological engineering, scientific knowledge management.

1. Introduction

Knowledge management has become recently a hot topic not only in many research communities all over the world but also increasing interest may be noticed in market companies getting lost in their own information sources of different kind. At the same time arising concept of Business Intelligence 2.0 is moving towards proactive approach to problem solving in business environments, instead of reactive, latter being implemented as searching for patterns in collected information to improve future decisions (knowledge discovery and data mining). In order to act before an event occurs, one must have a rapid access to the sources of knowledge critical to his decisions. Not only written documents, but also multimedia content [1] and domain experts are the subjects of knowledge management as important explicit and tacit knowledge resources. Our experiences with work in a big telecommunicatins market company, shows that the issues of knowledge management are in a very immature stage in there, but at the same time there is quite a big need to implement some adequate solutions in the area.

Even more complicated sounds a question of how to manage knowledge in a research company or institute, where problems being solved are usually more complicated than those in the commercial environments. Moreover, the final product of a research institute is the knowledge itself. Thus there is a strong need to organize and support development of creative environments [2] to improve the quality of research.

Ontologies as a knowledge representation method became very popular[1] in recent years, especially in computer science community after popularizing the idea of semantic web [3] as the future of the Internet. Number of standards, tools and languages supporting the idea has been developed since then. But on the other hand, it is hard to name a set of mature, well-established and widely used software implementations of knowledge management for market or research use[2].

Ambition of our research group in National Institute of Telecommunications (NIT) is to create the research institute knowledge management system (RIKMS) fit to the needs of a research institute, based on ontology engineering tools and methods and employing the ideas of creative space described in [2], [4].

This preliminary paper is mainly devoted to the problems of generation and maintenance of research topic ontology and is structured as follows. Section 2 discusses differences between two main ontological views on activities of a research institute. Some remarks on the way an ontology of research topics shall be structured are presented in Section 3. Section 4 is devoted to the general framework proposed for ontology creation and maintenance while Section 5 summarizes the paper and gives some directions of future development.

2. Organizational and Topic Ontology

The RIKMS reflects two different, but interrelated, perspectives on the knowledge maintained in a institution. The former is concentrated on the organizational aspects or on how to organize knowledge intensive processes and the latter is focused on research topics lying in the field of interest of the institution, or on how to refer to research areas and topics.

Building blocks of the organizational ontology reflect the structure of an organization, its working regime, accepted standards, policies and procedures, worked out prod-

[1] At least judging from the number of books and publications in this area.

[2] Contrarily to, e.g., knowledge discovery and data mining tools.

ucts, etc. Organizational ontology is usually hardwired in a software system dealing with its concepts, taking form of relational tables with fixed structure and hard-coded procedures expressed in some chosen language. On the other pole would be a software system storing and processing organizational ontologies in more general form as constructs of one of the ontology engineering languages[3]. However this is a completely different subject lying rather in the scope of software engineering methods and will not be discussed here.

By topic ontology we mean here a set of interrelated topics researched by the institute. It creates a different point of view on the activities performed within the institute. Organizational ontology may be seen as orthogonal to topic one and consequently they intersect each other. Projects may be indexed with keywords taken from the topic ontology of the institute, similarly employees will manifest competences in some topic labelled research areas.

Research and scientific institutions usually tend to form a hierarchical organizational structure constituted of departments, divisions, laboratories or working groups, centered around some research issues. Every unit of an institute has its own leader and employs people with similar educational and scientific background. Furthermore intellectual heritage and common sense of every group of this kind has been formed by its history, tradition, shared values, cooperation with external partners, long-term project experience. Such a group is thus thematically homogeneous to some extend and hermeneutic horizons[4] of its members are more coherent with respect to one another than to members of the other organizational units.

Reverse influence of research topics on organizational structure may be observed as well when we consider the origins of units within an institute. They are often formed around charismatic leader, transformed from successful working groups or answer a need to undertake a research in some previously uncovered area.

Summing up organizational and topic ontologies are closely bounded and one cannot drop any of them when dealing with the subject of knowledge management in a research institute. Our approach to ontology construction utilizes organizational structure as a framework for topic ontology creation and maintenance.

3. Topic Ontology Representation for a Research Institute

There are several motivations for creating a topic ontology for a research institute. Ability of viewing processes and their outcomes from the perspectives of projects, their products or people involved in them is attainable as all they are distinguishable concepts of organizational ontology and thus might be somehow reflected in the structure of knowledge management software system. But the questions immediately arise of how an overview of the activities from the research topics perspectives may be achieved or what the set of all topics researched by the institute consists of. Possibility of taking topic centered perspectives on projects, products, employees and documents has a meaningful importance for people involved in management of a research institute and heads of its departments. It supports many decision making tasks. Lets enumerate some of them.

- Reporting the achievements in particular fields of scientific activity entails a reflection on appropriateness and up-to-dateness of current organizational structure and enables build-up of development strategies.
- When applying for a new project or analyzing research trends, topic ontology centered view helps to determine whether an institute has enough expertize in related thematic fields.
- Knowing the competences of individual employees is a key prerequisite for building up interdisciplinary working groups capable of dealing with complex problem with many diverse research threads.
- Analysis of structure of topic ontology may lead to identifying the germs of new research topics.

On the other hand, topic ontologies may be useful at the individual researcher level as the important input for the tools supporting creativity. The idea of hermeneutic EAIR (enlightenment, analysis, hermeneutic immersion, reflection) spiral of searching through rational heritage of humanity and reflecting on the object of study has been presented in [4] with experiments on ontology supported hermeneutic agent, helping a user in search for knowledge sources related to object under research, reported in [2]. The ontology is used there to define researcher semantic profile that machine is able to process and use in order to help in finding relevant knowledge resources on the world wide web.

Textual and multimedia information is not the only source of knowledge in the research institute. Having an access to semantic profiles of institute employees, software agent might locate a person with strong competences in the subject of study. It could be a hard task for someone not familiar with everyone's research interests, but a computer fed with profiles of individuals can be very helpful.

3.1. Local versus Global Ontology

Different applications of topic ontology demand different views on set of concepts and relations. Intuitively, at the individual level, granularity is to be greater and ontology more detailed, as it supports actions performed during everyday work, at rather operational level. Higher in the or-

[3] Many standards have been developed. Let us mention web ontology language (OWL), or lower-level resource description framework (RDF).

[4] Hermeneutic horizon following H.-G. Gadamer [5] is "*The totality of all that can be realized or thought about by a person at a given time in history and in a particular culture*".

ganizational hierarchy of the institute, more general views on topics are needed as the horizon of group activities is more strategic.

Not every concept and relation is meant to be visible at the higher levels, some of them may remain private, but those of higher levels must be more reliable, commonly agreed and formal.

Distinctions mentioned above, along with general remarks on orthogonality of organizational and topic ontologies (see Section 2) lead to conclusion that it seems to be more reasonable to maintain distributed ontologies associated with different levels of organizational hierarchy, from individual, through group, ending up with an overall ontology of the institute.

As the responsibility for communication of an organizational unit with its environment lays on unit or, more generally, group leader, his or her role in ontology construction and maintenance processes should be superior. The leader is to be especially involved in the mechanism of keeping ontology of his group consistent and integrated with those of higher levels. We shall emphasize the role of the leader in proposed framework.

What must be decided next is whether there should be one ontology defined globally for the whole domain[5] and then adapted by its constitutive units or the better solution is rather to develop local ontologies for all individuals and combine them into higher level ontologies. We believe that the bottom-up strategy is a better solution. The intuitive justifying argument is that in the top-down procedure, there must be overall ontology defined, detailed enough to help in creativity support processes and, at the same time, covering all possible topics lying in the field of interest of the institute. Lets assume, we wish to adapt some kind of telecommunications ontology defined by ITU[6]. Saying that, e.g., National Institute of Telecommunications covers the whole universe of telecommunication related issues as defined by ITU, and nothing more, is not neccessarily a true statement[7]. From the one point of view that would be a nice property as it could enable adaptation of single, global view on activities of all telecommunication institutes all over the world. But at the same time it introduces informational mess, by importing to institute's research field concepts that are out of its scope and forgetting those which are applicable. The question of topic map for the whole institute would remain unanswered.

Attemps to build up a NIT ontology ([6] and further work), showed that the institute is active in a variety of diverse and advanced research fields, including typical low and high level telecommunications problems like electromagnetic compatibility, radiocommunications and mobile telephony, optoelectronics, network infrastructure and management, but also law, social and market issues like regulatory problems, customer satisfaction surveys and decision support problems including knowledge discovery and management, game theory and logic.

Having above considerations in mind, it seems to be much more promising approach to build an ontology starting from individual level, promote their local concepts to higher levels of department and institute in some manner and integrate them to achieve the global picture of institute activities. Such an idea of heterogeneous ontologies in distributed environment has been discussed in [7]–[9].

We stated above that every individual and group hold their own ontology. Group has been defined as an organizational unit, like department or laboratory. However within an institute there may exist a number of task teams, interdisciplinary groups with people primarily affiliated with different organizational units. Research institute as the whole may be seen as a group too. All those meanings of a group should be enabled to have their own common ontologies. In such a context, group leader is a head of an unit, or director of the institute, but also informal group leader or a person designated to take care of public image of, and knowledge management in his or her group.

3.2. Light versus Heavyweight Formalism

The word ontology has been being present in the common vocabulary for several centuries. Following is the Wikipedia's definition of its philosophical connotations:

> ... the oldest extant record of the word itself is the Latin form ontologia, which appeared in 1606, in the work Ogdoas Scholastica by Jacob Lorhard (Lorhardus) and in 1613 in the Lexicon philosophicum by Rudolf Göckel (Goclenius). The first occurrence in English of "ontology" as recorded by the OED appears in Bailey's dictionary of 1721, which defines ontology as "an Account of being in the Abstract".

Next search on the word ontology in Wikipedia, but this time in the context of computer science, gives next definition:

> ... an ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts.

The key word above is "formal". Increasing the amount of formalism gives many different realizations of ontology. Classification framework has been proposed in [10], while referring paper [11] discusses, also interesting from our point of view, issues of ontology based documents annotation. Lets enumerate definitions taken from [10] in a light-heavy order:

controlled vocabulary – finite list of terms,

glossary – list of terms and their meanings,

thesauri – list of terms with synonym relationship defined (but not an explicit hierarchy),

is-a – hierarchy of classes and their instances,

[5] Here: research institute.

[6] International Telecommunication Union – United Nations agency for information and communication technologies.

[7] Where is the place for this paper in such a case?

frames – classes having properties (including relations to the other classes),

value restrictions – characteristics of properties (e.g., type/class, value) restricted,

disjointness, inverse, part-of – additional relations between classes with well-defined semantics,

general logical constraints – arbitrary logical statements on classes, properties and instances.

Most of people require at least "is-a" level definition to be achieved to consider specification of the domain to be an ontology.

In our case it is still unclear how formal the model will be. We initially assume it to be at "is-a" level with possibility of defining additional, other than "is-a", somehow restricted, relations between classes. Nevertheless this setup may be significantly modified while the system evolves.

4. Topic Ontology Creation and Maintenance Process in a Research Institute

In this section a general framework of the process of ontology creation and maintenance for a research institute is presented. Must be stated that the work documented below is at the early stage and only limited number of details may be provided apart from the general idea.

In the following, for the sake of brevity, a set of interrelated ontological concepts associated with person, unit or institute will be called an ontological or semantic profile or simply a profile.

The basic scenario for ontology construction starts with submission of the new knowledge resource to the RIKMS. It is then analyzed by some automatic concept extraction method from text[8].

Then phase of individual[9] reflection is initialized, with querying an user on a relevance of discovered topics to his or her individual profile, their relations to existing topics in the profile and profiles of other people or higher level profiles and thus stimulate user to take a reflection on the profile and modify it accordingly. This step may be viewed as a limited form of analysis transition in EAIR spiral [4] as a localization of the new concepts in the context of existing semantic profiles.

Submitted document will be then indexed with new topics signatures and may serve as a proof of user's competence in the field characterized by those topics. Moreover, if submission occurred in some specific context, as a final report from project or part of some other activity, then system stores the relation of topics, document and context for further use.

Final cross-level agreement stage starts with identifying those parts of ontology which could be promoted to one of the higher levels of organizational hierarchy. During the process of system guided debate between all interested parties, concepts and relations are delegated up the hierarchy on the basis of common agreement, but with the deciding vote of a group leader. Cross-level agreement stage in the context of knowledge creation theories may be seen as a counterpart of debating transition of EDIS (enlightenment, debate, immersion, selection) spiral [4].

4.1. Topic Ontology Generation

This section describes document-based concept extraction method, being automatic step of ontology construction process. Different methods of automatic ontology construction have been surveyed in [12]–[14]. The approach proposed in this paper is similar to those used in On-To-Knowledge project for retrieval of relations between concepts from documents [15], [16].

In [6] a number of tools for automatic concept extraction have been outlined and the results of experiments conducted on publications from *International Journal for Telecommunications and Information Society* have been presented. Special attention has been paid to OntoGen system [17]. The reader is encouraged to refer to [6] for more details.

OntoGen generates an ontology in a semi-automatic process on the basis of a corpus of documents describing specific domain or subdomain of interest. Document clustering based method is used which needs some minimal number of documents to be available. It is rather designed for attaining a global view of a domain under investigation.

Chosen for our framework topic generation routine utilizes the well-known idea of frequent itemsets discovery in transactional data [18].

In preliminary step document is being transformed into transactional data by decomposing it into a set of sets of words (wordsets). Each wordset is roughly equivalent to a sentence in a grammatical sense.

Frequent wordset is defined as follows. Having a set of wordsets $D = \{S_i\}_{i=1,..,N}$ frequent wordset is a set of words $S^* = \{w_k\}_{k=1,...,M}$ simultaneously contained in at least $Supp\%$ of wordsets from D, where $Supp$ is called a support of a wordset:

$$S^* \text{ is frequent } \iff \frac{|\{S_i \in D \mid S_i \supseteq S^*\}|}{|D|} \geq Supp\,.$$

The motivation behind searching for frequent wordsets in a document is rather straightforward. If the same group of words has been used by authors in a number of sentences it may designate a concept or a set of related concepts.

After frequent wordsets are discovered, second operation is performed, so called pruning step. It is obvious that if a wordset S_i^* is frequent then every wordset S_j^* being a subset of S_i^* is frequent as well, as it is at least contained in the same set of sentences as S_i^* is. For the sake of brevity, all wordsets being subsets of frequent wordsets are removed

[8] At the moment we assume knowledge resources to be textual documents.

[9] Local or in other words happening between software system and its user.

from the result in a pruning phase. All, but not those which exist in significantly greater fraction of sentences than their supersets and thus might indicate more general concept.

Algorithm 1: Frequent word sets mining for ontology construction

```
Data: document
Result: nondominated_frequent_groups
norm_document = ConvertAndPreprocess(document)
transactions = MakeTransactions(norm_document)
frequent_groups = Apriori(transactions)
nondominated_frequent_groups = Prune(frequent_groups)
```

Table 1
Example of frequent word groups

Words	Count
Paper I	
theory, rational, intuition	16
normal, creation, knowledge	14
spiral, creation, knowledge	13
creation, knowledge	51
seci, spiral	19
humanity, heritage	16
space, creative	16
civilization, knowledge	15
dimensions, creative	14
shinayakana, systems	13
approach, systems	13
tacit, knowledge	13
heritage, knowledge	13
Paper II	
triple, helix, creation, knowledge	6
processes, creation, knowledge	10
indicators, quality, reference	9
spirals, creation, knowledge	8
normal, creation, knowledge	7
academic, creation, knowledge	7
spiral, creation, knowledge	7
minimized, maximized, indicators	6
units, sets, data	6
maximized, indicators, quality	6
best, sets, data	6
sets, data	14
profile, reference	12
Paper III	
nakamori, wierzbicki, spiral	10
academic, creation, knowledge	9
technology, management, knowledge	9
science, systems, knowledge	9
science, management, knowledge	9
creation, knowledge	34
management, knowledge	27
science, knowledge	18
nanatsudaki, model	16
seci, spirala	12
processes, knowledge	12
academic, spirals	10
processes, creation, knowledge	10
Paper IV	
representation, multiple, aggregation, criteria, knowledge	13
coefficients, weighting	19
integration, knowledge	15
form, knowledge	15
sum, weighted	14
reservation, aspiration	13
aspiration, levels	13

The four steps procedure starts with the preparation of document to make it fit to input requirements of the word groups searching routine. Format conversion to plain text, font encoding translation, removal of stop words, lowercasing and stemming are the main steps in preprocessing phase. After accomplishing that part, document is transformed into transactions that are fed to a frequent wordsets discovering algorithm. Finishing pruning step reduces the number of word groups and gives the final result.

Table 1 shows frequent word groups with cardinality greater than two for some papers thematically located in the field of knowledge science.

4.2. Individual Reflection

The purpose of the next stage is to populate individual profile with newly discovered topics and to define relations to the concepts found in both individual and neighbourhood profiles. It takes form of interaction between user and software system driven by word groups discovered in previous step and current structure of profiles. Sovereignty of the user is a superior principle. He or she decides on the final shape of individual profile. The role of software system is in stimulation of user's reflection by requesting advice about a local hierarchy of ontology. It might be seen as an engine searching for new topics and relations in both new source of knowledge and already established semantic structures. The final decision is always left to the user.

System is detecting whether any of basic indicators of new concept or relation existence arise and should be reported to the user providing a new source of knowledge.

Frequent word groups. Automatic topic generation phase proposes a number of words – candidates for new topics and indicates their coincidence in contexts of sentences. The n-grams for $n > 1$ are expected to be more informative than unigrams. They may carry two meanings. One is that there exists a concept in the domain identified by the name consisting of more than a single word. Second interpretation is a set of related concepts. Mix of those two is possible as well. Such a hypothesis of distributional semantics[10] lies in the basis of, e.g., some of the text summarization systems [19]. Scenario for this step is to present frequent groups to the user and let him decide whether they can contribute to his individual semantic profile and if so then whether they form single concept or group of interrelated topics. This may be enhanced by presenting some additional information like showing the quotations from the submitted document in which they appear.

Super- and sub-wordset. If user designate word groups being in a subset-superset relation, as bases for new concepts, it may be an indicator of existence of one of important relations "is-a" or "part-of" between corresponding concepts. System should suggest such a solution.

[10] Context has a strong influence on the word meaning.

Integration with existing profiles. New topics and relations evolve in the semantic context of person and neighbourhood. Thus during questioning process, system should take into consideration existing concepts and ask the user to localize newly discovered ones in the whole semantic profile. Searching for counterparts of ontological concepts is a subject of research in the area of ontology matching and alignment [20]–[22]. Variety of techniques are available from simple matching by name ending up with more sophisticated methods. This issue still needs to be investigated in more detail.

User invention. System should be able to process any other modifications proposed by the user at this stage. However some kind of constraint for user's freedom should be applied. It might be an obligation to provide an explanation of why the modification had been made. System may ask for a reference to a source of information on the new topic as a kind of evidence or to place new elements in the current ontological structure by linking them to existing topics.

4.3. Cross-Level Agreement

As mentioned in Subsection 3.1 the framework we are aiming at shall generate global view of the topics researched in the institute by aggregating individual profiles of employees. Special role is granted to a group leader, who has a final deciding vote as a person responsible for the overall picture of activities performed by his or her group.

Software system is engaged in two aspects of creating an agreement on ontology structure. First, it again stimulates a reflection on how the higher level structure should look like, by informing individual users on existence of potentially promotable concepts and relations. A number of indicators might be considered. Below a couple of exemplary ones are listed.

Shared concept. Sharing a concept between two or more individual profiles seems to be a good reason to promote it to the higher level. Both profile holders should be notified about the match found and decide together whether publish the concept or not.

Shared relation. One meaning of relations sharing is analogous to concept sharing. The second one is that relation linking concepts that had been promoted to the higher level should get high score as well.

Superconcept. Some of the relations are distinguished among others. For instance "is-a" associating super- and subconcept play a special role in any system as it introduces a hierarchy into it. Therefore superconcept of a concept promoted to the higher level should get a high score.

Strongly supported by the sources. Concepts from a individual profile having many knowledge sources associated are more likely to be promoted.

Existing ontology. Relations and concepts imported from another, especially widely recognized ontology are desirable. However they first should appear in at least one individual profile to justify their relevance to the institute.

User invention. Again system should give its users freedom to promote concepts and relations they consider to be important.

The second task for software system within a process of achieving cross-level agreement is assistance in debating on the shape of higher level structure. After a part of individual ontology is proposed to be promoted by the user, system should notify group leader and provide him or her with all acquired information about new ontological findings. After the final decision is made system should propagate updated ontology among all parties that may be interested in it and, in case of need, initiate a debate leading to the agreement.

5. Conclusions and Future Work

In this paper the framework for ontology construction for a research institute has been proposed. The framework is organized in a distributed hierarchical structure, with local ontologies associated with individual employees and an integrated higher level group ontologies with concepts and relations promoted from individual profiles. Three main steps of ontology construction have been outlined, namely topic generation from documents, individual reflection on ontological profile and cross-level agreement between interested parties. Special, superior role of group leader has been emphasized. Some preliminary results for simple, but robust topic generation method have been presented.

We believe the framework may be a better choice for a research institute trying to develop its own ontology of research topics for integration and management of knowledge resources than adaptation of any well-known domain ontology or creation of global ontology by domain experts. Reflection and agreement stages have themselves an additional value as they are driving processes of exploring scientific neighbourhood (individual reflection) and exchanging knowledge through debate (cross-level agreement). As such they may be seen as supporting creativity in scientific environment.

What must be stressed here is that development is at the early stage and far from complete. There is still much of work to be done. More sophisticated methods for topic extraction from documents are to be tested, detailed specification of reflection and agreement phases and implementation of software component with appropriate human-computer interface is still to be worked out.

References

[1] *Semantic Multimedia and Ontologies. Theory and Applications*, Y. Kompatsiaris and P. Hobson, Eds. New York: Springer, 2008.

[2] *Creative Environments: Issues of Creativity Support for the Knowledge Civilization Age*, A. P. Wierzbicki and Y. Nakamori, Eds. *Studies in Computational Intelligence*. Berlin: Springer, 2007, vol. 59.

[3] T. Berners-Lee, *Weaving the Web: the Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco: HarperSanFrancisco, 1999.

[4] *Creative Space: Models of Creative Processes for the Knowledge Civilization Age*, A. P. Wierzbicki and Y. Nakamori, Eds. *Studies in Computational Intelligence*. Berlin: Springer, 2006, vol. 10.

[5] H.-G. Gadamer, *Truth and Method*. 2 ed. New York: Crossroad, 1989.

[6] A. P. Wierzbicki and E. Klimasara, "Wybrane problemy zarządzania wiedzą. Zadanie 1: Praktyczne aspekty zarządzania wiedzą", Statute Project nr 06300017 Internal Report. Warsaw: National Institute of Telecommunications, 2007 (in Polish).

[7] M. Uschold, "Creating, integrating and maintaining local and global ontologies", in *Proc. First Worksh. Ontol. Learn. OL-2000 14th Eur. Conf. Artif. Intell. ECAI-2000*, Berlin, Germany, 2000.

[8] P. R. S. Visser and Z. Cui, "On accepting heterogeneous ontologies in distributed architectures", in *ECAI'98 Worksh. Appl. Ontol. Probl.-Solv. Meth.*, Brighton, UK, 1998.

[9] A. Kuczynski, D. Stokic, and U. Kirchhoff, "Set-up and maintenance of ontologies for innovation support in extended enterprises", *Int. J. Adv. Manuf. Technol.*, vol. 29, no. 3/4, pp. 398–407, 2006.

[10] O. Lassila and D. L. McGuinness, "The role of frame-based representation on the semantic web", Technical Report 01-02, Knowledge Systems Laboratory, Stanford University, 2001.

[11] O. Corcho, "Ontology based document annotation: trends and open research problems", *Int. J. Metadata Semant. Ontol.*, vol. 1, no. 1, pp. 47–57, 2006.

[12] "A survey of ontology learning methods and techniques", A. Gómez-Pérez and D. Manzano-Macho, Eds. OntoWeb Deliverable D1.5, Universidad Politécnica de Madrid, 2003.

[13] P. Buitelaar, P. Cimiano, and B. Magnini, *Ontology Learning from Text: An Overview*. *Frontiers in Artificial Intelligence and Applications Series*. Amsterdam: IOS Press, 2005, vol. 123.

[14] C. Biemann, "Ontology learning from text: a survey of methods", *LDV-Forum*, vol. 20, no. 2, pp. 75–93, 2005.

[15] J. U. Kietz, R. Volz, and A. Maedche, "A method for semi-automatic ontology acquisition from a corporate intranet", in *Proceedings EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France, October 2000*, *LNAI*, vol. 1937. Berlin: Springer, 2000.

[16] J. U. Kietz, R. Volz, and A. Maedche, "Extracting a domain-specific ontology from a corporate intranet", in *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop, Lisbon, 2000*, C. Cardie, W. Daelemans, and C. Nédellec, Eds. New Jersey: Association for Computational Linguistics, 2000, pp. 167–175.

[17] B. Fortuna, M. Grobelnik, and D. Mladenic, "Semi-automatic data-driven ontology construction system", in *Proc. 9th Int. Multi-Conf. Inform. Soc.*, Ljubljana, Slovenia, 2006.

[18] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases", in *Proc. 1993 ACM SIGMOD Int. Conf. Manage. Data*, P. Buneman and S. Jajodia, Eds. Washington, USA, 1993, pp. 207–216.

[19] C.-Y. Lin and E. H. Hovy, "The automated acquisition of topic signatures for text summarization", in *Proceedings of the 18th International Conference on Computational Linguistics, Universität des Saarlandes, Saarbrücken, Germany*. San Francisco: Morgan Kaufmann, 2000, pp. 495–501.

[20] M. Ehrig, *Ontology Alignment: Bridging the Semantic Gap (Semantic Web and Beyond)*. New York: Springer, 2006.

[21] M. Ehrig and J. Euzenat, "State fo the art on ontology alignment", Knowledge Web Deliverable D2.2.3, University of Karlsruhe, 2004.

[22] N. T. Nguyen, *Advanced Methods for Inconsistent Knowledge Management (Advanced Information and Knowledge Processing)*. New York, Secaucus: Springer, 2007.

Cezary Chudzian received his M.Sc. in computer science from the Warsaw University of Technology, Poland, in 2002. He is a researcher at the National Institute of Telecommunications. Currently he works on his Ph.D. in the area of knowledge management. His main scientific interests include: practical applications of knowledge discovery techniques, machine learning theory, knowledge management, global optimization, and advanced software engineering.

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

Paper

Towards a Unified Architecture of Knowledge Management System for a Research Institute

Jarosław Sobieszek

Abstract—This paper presents some elements of architecture of planned knowledge management system dedicated to research institutions. Main contributions include social extension of the idea of adaptive hermeneutic agent and preliminary implementation of domain specific language for development of knowledge management systems. Work described here concentrated on practical verification of viability of proposed ideas and took form of a prototype software system, which can be used by a group of researchers to easily find and recommend relevant information.

Keywords—creativity support, knowledge management, model-driven software development, tagged collaborative filtering.

1. Introduction

One of the possible definitions of knowledge management [1] is:

"*Knowledge Management is the discipline of enabling individuals, teams and entire organisations to collectively and systematically create, share and apply knowledge, to better achieve their objectives*."

The tasks of knowledge management can be summarized by a checklist often used by journalists to verify that they present the whole picture of a situation. This list, known as *Five Ws and H*, consists of six interrogatives, which when applied to knowledge management [2] roughly correspond to:

- "when"' – time management,
- "what" – task management,
- "how" and "where" – information management,
- "who" – people management,
- "why" – goal management.

As can be seen, this list encompasses a wide range of different fields and potential techniques. Very often the solutions tend to be tailored to the needs of commercial entities, since they are created by large software corporations, which need to recoup their investment. The needs and expectations of research institutions are, to a significant degree, different and were only partially explored by theorists and especially by practitioners of knowledge management. One of the possible reasons for this disparity is a fact that processes of knowledge creation in academia are quite different to the ones at commercial institutions [3].

Our group at National Institute of Telecommunications, motivated to a significant degree by local practical need, decided to explore the topic of knowledge management in research institutions. Our main goal is creation of an integrated system which will merge traditional approaches to knowledge management with theories of creativity support [3]. Secondary requirements include architectural flexibility, which should simplify planned future deployments in other institutions, and low cost of proposed solution, to expand the group of potential users.

This paper presents results of an experiment conducted to investigate feasibility of proposed approach to development of knowledge management systems. We decided to limit the scope of this study to a small, but none the less useful part of the complete system, and focus our attention on the creativity support component. Section 2 presents three topics relevant to the presented application outlining, respectively, theory of knowledge creation, representation of preferences and interests, and integrated approach to development of information systems. Section 3 justifies some of design decisions and describes structure of the prototype. Section 4 summarizes main results presented in this article and details potential further enhancements and directions of future research.

2. Background

2.1. Creative Environment

Creative environment [3] is a comprehensive theory describing a place, where knowledge is created, shared and used. One of the most important aspects of this idea is identification of various knowledge creation processes and description of ways to support them.

The basis of this theory is formed by a model of knowledge creation called Nanatsudaki. The name is a Japanese phrase meaning *seven waterfalls*, and corresponds to the structure of this model. It is composed of seven so called knowledge creation spirals, which describe processes of knowledge creation typically encountered in both research and industrial institutions. The cyclic nature of the spirals reflects the fact, that knowledge creation is a perpetual and self-propelled (positive feedback) endeavour.

One of these processes, called hermeneutic spiral, describes the activity of gathering sources, most often in a form of publications by other researchers, analyzing them and reflecting on them in search of new research ideas. It forms the very basis of a large part of scientific work. The hermeneutic spiral is also known as EAIR (enlightenment-analysis-hermeneutic immersion-reflection) spiral, an acronym derived from the four phases of this knowledge creation process.

- Enlightenment is a phase that starts with having an idea, which is considered to be worthy of further involvement and during which potential sources of information are explored and research materials are gathered.
- Analysis is a phase of rational study of relevant materials.
- Hermeneutic immersion is a phase during which concepts and ideas rationally explored in previous stage are absorbed into one's intuitive perception.
- Reflection is a phase during which new research ideas are intuitively considered and explored.

Another related concept is that of adaptive hermeneutic agent (AHA), a software system designed to support the knowledge creation process represented by the EAIR spiral. The original idea [3] described a system, which helped individual researcher to find relevant materials on the web and which was largely based on algorithmic analysis of document content. We decided to replace this mechanisms with framework for cooperation, motivated to some degree by growing importance of social web sites. This approach encourages collaboration and will hopefully lead to a more comprehensive realization of the idea of creative environment.

2.2. Tagged Collaborative Filtering

Knowledge representation is another important aspect of the architecture of knowledge management system. Here, we describe the representation of preferences and interests, since this information forms the basis of the proposed system.

Nowadays, Internet stores routinely use so called recommender systems [4] which propose goods the customer might be interested in buying. Generally, approaches to construction of these systems fall into two broadly defined categories.

First of them, called *content based*, concentrates on creating profiles which aim to explicitly describe both users and products. This technique relies on additional domain specific information, which could be hard to gather.

Alternative approach called *collaborative filtering* [5] or social information filtering relies only on past user behavior, predicting future interest based on preferences that were expressed by a preferably large group of users. These preferences could have been specified explicitly, taking form of ratings which quantify level of satisfaction, or could be extracted from more implicit sources, such as histories of purchases or page views. Generally such information is more readily available which partially explains relative popularity of solutions based on collaborative filtering. Additionally, it is a more versatile approach, since it does not depend on content being recommended.

Formally, let U and I denote, respectively, sets of users and items, with $|U| = n_U$ and $|I| = n_I$. Rating function $r : U \times I \rightarrow S$ is a mapping of user-item pairs into a rating scale S which is most often represented as a sequence of natural numbers, usually of length 5 or 10. Values of this function for a given sets of users and items can be tabularized to form a matrix $R = [r_{ui}]_{U \times I}$, where r_{ui} is a rating given by user u to item i. This formulation of the problem allows us to alternatively define collaborative filtering (or at least its most common form) to be an algorithm for estimation of missing entries of a matrix.

Tagging is another method commonly used for knowledge representation. The idea is to associate short phrases, known as tags, to provide additional information about some data. This is closely related to a concept of keywords used by librarians to index textual resources.

This two approaches can be merged to form what we have called *tagged collaborative filtering* which can be viewed as a multicriteria variant of collaborative filtering. Standard formulation of multicriteria analysis requires all values of the criteria to be specified, so a different name better reflects the fact that in this case they are optional. This approach is also quite similar to some of the methods used for content based recommendation systems, though one significant difference is that the tags can be used to not only describe content, but also, for example, preferences of the users.

Formally, we introduce another dimension into domain of the rating function which now becomes $r : U \times I \times T \rightarrow S$, where T is a set of tags. Both of the constituent ideas can now be expressed by imposing some limits on the dimensionality of sets used in its definition. Tagging is equivalent to reduction of rating scale S to a binary alternative, collaborative filtering is equivalent to reduction of tag set T to a single implied value which can be called quality or satisfaction.

Since collaborative tagging can be viewed as an approach to ontology construction, it should be possible to further extend this idea, and apply more sophisticated semantic structures to describe relations between tags, which could be then used for collaborative filtering.

2.3. Model-Driven Software Engineering

Software development is a complex process. One of the most common approaches to dealing with complexity is an idea of splitting the problem into parts and dealing with them on an individual basis, also known as *divide and conquer*. When it is applied on a conceptual level, the parts are often called layers. This approach, when applied to software engineering, usually splits the process into three phases (analysis, design and development) corresponding to semantic, structural and technological aspects of the problem.

Duplication is probably one of the most pervasive problems in software development. It is usually considered harmful to the quality of the affected systems, though there are some specific cases when it is actually helpful, e.g., loop unrolling which repeats statements in the body of the loop to reduce the number of tests and jumps, often leads to a faster execution of the program and is one of the techniques used for code optimization. The advice of avoiding duplication was expressed by a pragmatic rule of software development [6], known as DRY (don't repeat yourself).

The risks of duplication of information were recognized, for example, in a field of relational databases, where a design technique called normalization [7] aims to minimize structural problems associated with having multiple sources of the same data. Designs which do not follow this practice are more susceptible to the occurrence of so called data anomalies, which can lead to a loss of data integrity.

Canonical layered approach to software development does not have any mechanisms which prevent duplication. It can be seen as one of the tasks of project manager. This arrangement can fail, especially since higher layers of this process often produce only design documents, which are often perceived only as a direction for future work. Additionally, since lower levels build upon previous steps, they tend to rephrase at least some of the work that was already done, which can introduce inconsistencies.

Model-driven software development (MDSD) is one of the possible techniques, which help to reduce duplication. It is a design philosophy emphasizing the role of models as a cornerstone of process of software creation.

Structure of model is determined by another model, called metamodel, which can be seen as a specification of vocabulary that can be used to define models. This class-instance relationship can be extended indefinitely, though in practice there is usually no need to go beyond three levels, with the most generic one defined in a recursive way. Model level is application specific, metamodel level provides a generalized view of a problem domain, and metametamodel level is associated with software development environment allowing it to access lower level constructs in a standardized way.

Individual models can be connected with transformations (see Fig. 1), which describe methods of converting one model into the other. Usually, conversion of models to/from their textual representation is treated separately using techniques, which facilitate text parsing and generation.

The process of software development can be seen from a global perspective as a directed graph, whose nodes are models and edges are model transformations. The sources, that is the nodes which are not a destination of any transformation, represent models which need to be specified by the developer. The transitions are another part which needs to be defined. The output of the process is represented by sinks, that is the nodes which are not a source of any transformation. They correspond for example to source code, documentation or user interface definitions.

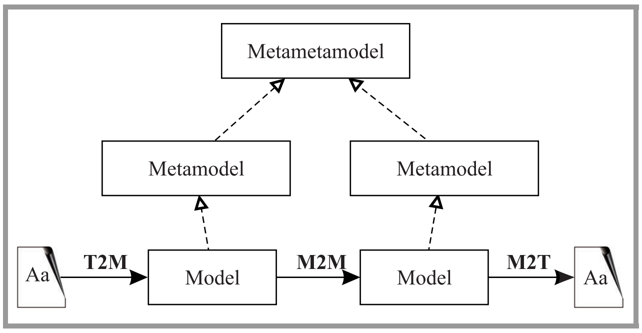


Fig. 1. Context of model transformations. Explanations: M2M – model to model, M2T – model to text, T2M – text to model.

One important consequence of imposing this kind of structure is that, the process of software development can be easily split into parts, which reflect certain perspectives, or ways of looking at the resulting system. For example, the process of development of data warehouse, can be split into several pieces: one that defines a transformation of domain specific model into a domain independent representation, the other one describes a way of implementing that representation in a specific runtime environment and yet another one specifies configuration information. This decomposition can reflect the structure of the development team, when the first transformation is defined by a business analyst, the second one by a software engineer, and the last one by maintenance staff.

Some of the other advantages of this approach include formalization of knowledge and greater potential for reusability. It forces the developer to formalize the approach used to solve the problem. From a point of view of future maintenance of the system it is a great advantage, since it documents all the decisions made by the developer and bridges the semantic gap that often arises between concept representations at different levels.

This technique is foremost a way to introduce static structure to the problem, so it won't be of much use in situations where the complexity is mostly of algorithmic nature. It will be of great help mostly in large heterogeneous information systems characterized by high structural and low algorithmic complexity, such as data warehouses or knowledge management systems.

We have decided to use probably the most popular approach to model-driven software development, namely model-driven architecture (MDA) [8]. It was developed under the auspices of Object Management Group (OMG), a widely known organization, which has, for example, standardized the Unified Modeling Language (UML).

The specific tool we have used is known as openArchitectureWare (OAW). It's a modular code generation framework, nicely integrated with Eclipse development environment and based on Eclipse Modeling Framework (EMF). One of the distinguishing features of this tool is its support for text to model transformations, which enables the developer to easily define domain specific languages (DSL).

3. Prototype

Primary goal of the work presented here was to explore the ideas and techniques described in Section 2. This examination took a form of a prototype software system, whose primary function is the ability to catalogue and search for various objects related to the field of research. On one hand, it can be viewed as a greatly simplified knowledge management system dedicated to research institutions, on the other hand, it is a social incarnation of an adaptive hermeneutic agent.

Social aspect of the system is emphasized by its approach to editing the data. It mimics wiki-like systems in that regard, allowing any user to add, edit and delete content from the database. With this freedom comes the disadvantage of increase in maintenance work, since information stored in the system needs to be protected from willful destruction. On the other hand, it lowers barriers to participation expanding the potential group of contributors. Wikipedia is a proof that this approach is both feasible and has a lot of potential.

Since semantic profile information needs to be stored on a per-user basis, to fully use the system one has to create an user account. The need to do this can be viewed as cumbersome, and potentially discourage some of the likely users. Therefore, we decided to make the registration process optional, and allow users to use the system without providing any additional information. Such passive users do not contribute to collaborative filtering, though hopefully if they find it useful, they will become more active participants. This reflects our philosophy that it is better to encourage than to force.

Another aspect that emphasizes this laissez-faire user experience is approach to ontology creation. Basically, there are two generic ways of building ontologies, known, respectively, as top-down and bottom-up approach. First of them is a more formalized process, where a group of experts progressively specializes the vocabulary used to describe the problem domain. Somewhat similar technique, known as mind mapping, is often utilized for brainstorming and note taking. The other approach starts with a collection of items describing the problem domain. They are analyzed to extract the most specialized concepts, which are then repeatedly generalized. This approach is susceptible to automation, where first step can use keyword extraction algorithms, followed by a series of clusterizations, to form the final ontology.

Ours is basically a bottom-up approach, though with one crucial difference, when compared to automatic method described above. It replaces computer algorithms with a framework for cooperation, which should allow interested parties to form the ontology as a byproduct of their evaluation of source material. This approach is known as folksonomy, which is portmanteau made by combining folk and taxonomy, and is often used to describe the emergent process of ontology creation happening in a group of collaborating people.

As was already mentioned, we decided to investigate the feasibility of using model-driven approach to construction of knowledge management systems. Thus, the backbone of prototype presented here is formed by a definition of a metamodel (Fig. 2), which formalizes vocabulary used to describe the structure of this system. It is a simple object-oriented representation, composed of classes, which besides having attributes for storing values, can also be connected to each other with one of the two relations, namely many-to-one and many-to-many. Additionally both system and classes definitions can be annotated with metadata, which are called options here, that have a textual form and were used to specify labels displayed in the user interface. While not very elaborate, this metamodel is sufficient to describe a wide range of practical applications.

```
System:
      "system" ":"
      (options+=Option | classes+=Class)*;
Option:
      "option" name=ID "=" value=STRING;
Class:
      "class" name=ID ":"
     (options+=Option | attributes+=Attribute)+;
Attribute:
      name=ID ":" type=Type (options=TypeParams)?;
Enum Type:
      string="String" | m2o="ManyToOne" |
      m2m="ManyToMany";
TypeParams:
       "(" TypeParam ("," TypeParam)* ")";
TypeParam:
      ID | INT;
```

Fig. 2. Specification of model parser.

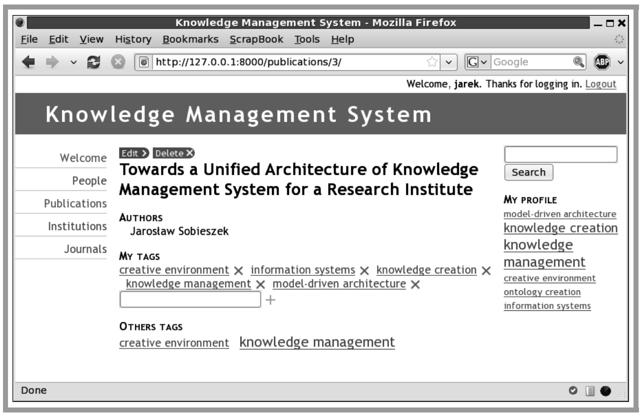
Based on the metamodel definition, we constructed a simplified model (Fig. 3) of publications catalogue. It consists of four classes, which represent respectively person, publication, institution and journal, connected with some self-explanatory relations. Thorough description of this particular application was not our goal, but it is something, that can be easily achieved. Thanks to chosen approach, what needs to be done from a technical point of view is a simple change of model definition. It is also possible to completely change the focus, and create, for example, a social bookmarking application or a movie database.

Again, all that is strictly necessary is a change of model definition.

```
system:
      class Person:
           first_name:  String(100)
           last_name:  String(100)
           affiliation:  ManyToOne(Institution)
      class Publication:
           title:  String(200)
           authors:  ManyToMany(Person)
           journal:  ManyToOne(Journal)
      class Institution:
           name:  String(200)
      class Journal:
           name:  String(100)
```

Fig. 3. Model of the prototype system.

The prototype took a form of a web application developed using Django framework. This allows it to be used on a variety of platforms, including, for example, mobile phones. Basic functionality focuses on providing create-read-update-delete (CRUD) interface to a catalogue describing some objects. User interface (Fig. 4) follows a common three-pane design. The central one displays information about object or a list of objects, the left one allows browsing specific classes of objects, and the right one provides interface for searching the database.



http://127.0.0.1:8000/publications/3/

Welcome, **jarek**. Thanks for logging in.

Knowledge Management System

Towards a Unified Architecture of Knowledge Management System for a Research Institute

AUTHORS

Jarosław Sobieszek

MY TAGS

creative environment ✕ information systems ✕ knowledge creation ✕ knowledge management ✕ model-driven architecture ✕

OTHERS TAGS

creative environment knowledge management

MY PROFILE

model-driven architecture knowledge creation knowledge management creative environment ontology creation information systems

Fig. 4. User interface.

Functionality related to recommendation is at the moment limited to tagging. Every object can be annotated with keywords, which are then displayed in two separate lists. First of them shows tags of a logged in user, second one of all the other users aggregated to form a tag cloud. Keywords used by user to describe objects form a profile, also displayed as a tag cloud, which enables easy access to related content. Without logging in user cannot associate keywords with objects, and can only see a list of tags added by other people.

4. Conclusions and Future Work

In this paper we presented some elements of architecture of planned knowledge management system dedicated to research institutions. Main contributions include social extension of idea of adaptive hermeneutic agent and early stage of implementation of domain specific language for description of knowledge management systems. Work described here was preliminary, and its main goal was verification that proposed approach is viable direction of future efforts. The results of this feasibility study were encouraging, and we intend to build upon them in our forthcoming projects.

One of the more evident directions of future work is extension of adaptive hermeneutic agent component, which was only partially implemented. Especially, to fully utilize it, the profile needs to be directly editable and allow for more direct specification of preferences. Also metamodel, even though it is sufficient to describe a wide range of real world applications, needs to be extended, if it is to be used for construction of more comprehensive knowledge management applications. One simple, yet very powerful, addition would be introduction of processes [9], which are widely used for description of sequences of actions and, thus, well suited to support many of management tasks.

Other more long-term possibilities include addition of different algorithms for constructive manipulation of data gathered in presented system. For example, network structures could be analyzed, to compute impact factor of objects [10]. Similar approach is used by some search engines [11], and would extend scope of potential applications. Also interesting would be formalization of semantic structure of this system, built upon work done in fields of ontological engineering and semantic web [12], [13]. This would make the data amenable to more intricate automatic processing.

References

[1] R. Young, "Definition of knowledge management" [Online]. Available: http://www.knowledge-management-online.com/Definition-of- Knowledge-Management.html

[2] R. Young, "The future of knowledge management" [Online]. Available: http://knol.google.com/k/ron-young/the-future-of-knowledge-management/1emn5abyls393/4

[3] *Creative Environments: Issues of Creativity Support for the Knowledge Civilization Age*, A. P. Wierzbicki and Y. Nakamori, Eds., *Studies in Computational Intelligence*. Berlin-Heidelberg: Springer-Verlag, 2007, vol. 59.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[5] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry", *Commun. ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[6] A. Hunt and D. Thomas, *The Pragmatic Programmer: From Journeyman to Master*. Boston: Addison-Wesley, 1999.

[7] E. F. Codd, "A relational model of data for large shared data banks", *Commun. ACM*, vol. 13, no. 6, pp. 377–387, 1970.

[8] O. Pastor and J. C. Molina, *Model-driven Architecture in Practice: A Software Production Environment Based on Conceptual Modeling*. Berlin-Heidelberg: Springer-Verlag, 2007.

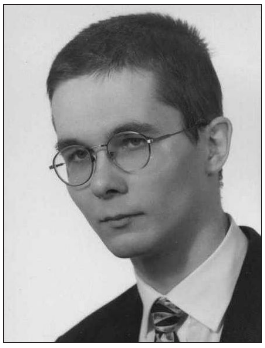
[9] M. Dumas, Wil M. van der Aalst, and A. H. ter Hofstede, *Process Aware Information Systems: Bridging People and Software Through Process Technology*. Hoboken: Wiley, 2005.

[10] J. E. Hirsch, "An index to quantify an individual's scientific research output", *Proc. Nat. Acad. Sci.*, vol. 102, pp. 165–169, 2005.

[11] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: the Science of Search Engine Rankings*. Princeton: Princeton University Press, 2006.

[12] G. Antoniou and F. van Harmelen, *A Semantic Web Primer*. Cambridge: The MIT Press, 2004.

[13] *Towards the Semantic Web: Ontology-driven Knowledge Management*, J. Davies, D. Fensel, and F. van Harmelen, Eds. Chichester: Wiley, 2003.



Jarosław Sobieszek received his M.Sc. degree in computer science from Warsaw University of Technology, Poland, in 2002. Currently he is a researcher at National Institute of Telecommunications, where he prepares his Ph.D. thesis in the area of knowledge management. His research interests include machine learning, artificial intelligence, knowledge management and model-based approaches to software development.

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

Paper

Multicommodity Auction Model for Indivisible Network Resource Allocation

Kamil Kołtyś, Piotr Pałka, Eugeniusz Toczyłowski, and Izabela Żółtowska

Abstract—In this paper we present the multicommodity auction model BCBT-I that allocates indivisible network resources among bidders. The approach can be considered as a generalization of the basic multicommodity model for balancing communication bandwidth trade (BCBT). The BCBT model assumes that offers concerning inter-node links and point-to-point bandwidth demands can be realized partially. However, in the real-world trade there might be a need to include capacity modularity in the market balancing process. Thus we state the model for balancing communication bandwidth trade that takes into account the indivisibility of traded bandwidth modules. This requires to solve a mixed integer problem and increases computational complexity. Furthermore, the pricing issue appears nontrivial, as the dual prices cannot be longer used to set fair, competitive market prices. For clearing the market, we examine the multicommodity pricing mechanism based on differentiation of buy and sell market prices.

Keywords— bandwidth allocation, indivisible commodities, modularity, multicommodity trade, pricing.

1. Introduction

In this paper the bandwidth trading is considered from the viewpoint of network operators, service providers and other wholesale active market players, buying and selling bandwidth. For the purpose of modeling trade of bandwidth resources in the communication networks, the network consists of nodes connected by links. The capacity of an inter-node link is the elementary commodity on the bandwidth market. However, network resources being traded can be more complex and can be composed of many links, i.e., paths, subnetworks.

It is well recognized that the base for bandwidth trading can be standardized contracts, that use prespecified amount of bandwidth [1]. This requires to take into account modularity of capacity in the trading models. Indivisibility can be associated with bandwidth sell offers concerning links or/and bandwidth buy offers concerning end-to-end network paths. Modularity requirements can be applied in trading resources of any layer of a communication network architecture (for example, optical links and synchronous digital hierarchy (SDH) containers). The size of indivisible unit of bandwidth may differ depending on the market considered or even depending on individual offers. For example, there may be portfolio of synchronous transport module (STM) contracts. Moreover, buyers may need to buy a set of different bandwidth links to establish connections. They should not be exposed to risk of buying some but not all links or risk of buying different quantities of bandwidth on different links.

Requirements of bandwidth market participants are difficult to satisfy using bilateral agreements, which are currently the most popular form of communication bandwidth trading. Other mechanisms, such as simple auctions and exchanges aim mainly in facilitating buyer-seller contacts. Thus, the efficient bandwidth trade requires development of advanced business tools [2], [3].

To cope with the problem of providing bidders with the possibility to submit offers for bundles of elementary commodities when auctioning indivisible units of bandwidth, researchers have proposed various rules and approaches. They can be assigned into two classes: simultaneous, single link auctions [2], [4]–[6]; and combinatorial auctions [7]. In simultaneous, separate auctions for individual links a user that wants to buy a certain path must put simultaneous bids at all relevant auctions. Then special, iterative mechanisms are required to coordinate individual links-auctions. This aspect, as well as possible suboptimality are the main roots of our criticisms for these methods. Combinatorial auctions based approaches may be seen as the best suited approaches for bandwidth trading. However they are proved to be NP-hard (non-deterministic polynomial-time hard) which is the main disadvantage pointed out also by other researchers. Lastly, all the over mentioned approaches require buyers to specify the particular links that constitute a desired path. This may lead to welfare inefficiency, as was shown in [8].

In this paper we state a multicommodity auction model BCBT-I that allocates indivisible network resources among bidders and provides efficient allocation of indivisible units of traded bandwidth resources. The model falls into a class of the multicommodity exchange models, that can provide efficient resources allocation by solving global economic surplus (welfare) maximization problem. The basic balancing communication bandwidth trade (BCBT) model proposed in [8] was the preliminary step in designing efficient multicommodity bandwidth exchange. The distinguishing feature of BCBT is that it allows bidders to place buy offers not for bundled links, but rather for end-to-end connections. This model is in the form of a linear programming problem in which many elementary buy and sell offers are simultaneously considered. Prices in BCBT model can be set according to values of appropriate dual prices. However, the basic BCBT model treats bandwidth as fully divisible commodity.

In the paper we provide a generalization of BCBT model, allowing us to consider capacity modularity requirements.

The issue is nontrivial, as the new model BCBT-I allows participants to declare different sizes of indivisible units of bandwidth to be traded. Moreover, the pricing issue appears as dual prices can not be longer used to set fair, competitive market prices. We examine application of multicommodity clearing mechanism based on differentiation of buy and sell competitive market prices.

The paper is organized as follows. The description of the BCBT-I model is given in Section 2. Section 3 discusses application of multicommodity balancing mechanism to fair distribution of social welfare. Section 4 presents simple examples illustrating the main features of BCBT-I model. Section 5 is the summary of the paper.

2. The BCBT-I Model

The BCBT-I model falls into a class of the multicommodity exchange models. It provides a considerable functional extension to the BCBT model [8], which treats bandwidth as fully divisible commodities. Buy and sell offers considered state capacity (appropriately demanded or supplied) and unit price (appropriately sell or buy). Realization of sell and buy offers is given by a non-negative variable less or equal to offered capacity.

Contrary, the BCBT-I model assumes that bandwidth is traded in indivisible amounts. Besides capacity and unit price, market participant may declare the size of indivisible unit of bandwidth in which the submitted offer should be realized, for example, 155.52 Mbit/s corresponding to one STM-1 contract. Such a feature may be very valuable in real trading practises, however it was not addressed by other researchers dealing with bandwidth indivisibility (see, for example, [5], [9]).

Market clearing with indivisible bandwidth requires to solve a mixed integer problem. This leads to integer variable problems, increasing computational complexity of the model comparing to simple BCBT model, as problem changes character from P to NP-hard. Thus applying BCBT-I for large networks with many market participants may require some aggregation mechanisms – an issue introduced in [10]. Below we give the statement of mixed integer formulation of BCBT-I model.

2.1. Mathematical Programming Formulation

The BCBT-I model defines three sets: network nodes (V), buy offers (D) and sell offers (E). Each buy offer $d \in D$ defines maximum capacity to be bought h_d, unit price E_d and size of indivisible unit in which bandwidth has to be purchased M_d. Each buy offer $d \in D$ concerns end-to-end path, described by the source node s_d and sink node t_d. Similarly, each sell offer $e \in E$ defines maximum capacity y_e, unit price S_e and size of indivisible unit in which bandwidth is offered for sale M_e. Sell offers concern particular links. This relationship is reflected by parameters a_{ve} defined for each pair $(v,e) \in V \times E$. Parameter a_{ve} accepts three values: 1 if a link connected with offer e originates in node v, –1 if a link connected with e terminates in node v and 0 otherwise.

It is assumed that offers can be realized in the indivisible units of bandwidth, thus x_d is the integer variable stating the number of units M_d realized for buy offer d, x_e is the integer variable stating the number of units M_e realized for sell offer e. Non-negative variable x_{ed} is continuous and denotes the bandwidth capacity allocated to sell offer e to serve buy offer d. The model BCBT-I is formulated as a mathematical linear program presented below:

$$\hat{Q} = \max\left(\sum_{d\in D} E_d M_d x_d - \sum_{e\in E} S_e M_e x_e\right), \quad (1)$$

$$0 \le M_d x_d \le h_d, \quad \forall_{d\in D}, \quad (2)$$

$$0 \le M_e x_e \le y_e, \quad \forall_{e\in E}, \quad (3)$$

$$\sum_{d\in D} x_{ed} \le M_e x_e, \quad \forall_{e\in E}, \quad (4)$$

$$0 \le x_{ed}, \quad \forall_{e\in E, d\in D}, \quad (5)$$

$$\sum_{e\in E} a_{ve} x_{ed} = \begin{cases} M_d x_d & v = s_d \\ 0 & v \neq s_d, t_d \\ -M_d x_d & v = t_d \end{cases}, \quad \forall_{v\in V, d\in D}, \quad (6)$$

$$x_d \in \mathbb{Z}, \quad \forall d \in D, \quad (7)$$

$$x_e \in \mathbb{Z}, \quad \forall e \in E. \quad (8)$$

The aim of the BCBT-I model is to maximize the economic welfare, which is the market surplus defined as a difference between buyers incomes and sellers costs – objective function (1). First and second group of constraints set upper and lower bounds on accepted volume of supply constr. (2) and demand constr. (3). Next two group of constraints ensure that total bandwidth flow at particular link will not be greater than realization of sell offer concerning this link constr. (4) and that bandwidth flow at all links will be non-negative constr. (5). Constraints (6) assert appropriate bandwidth flow for demand realization of each buy offer at each node and can be seen as an analogue to the Kirchhoff's current law. Two last groups of constraints impose indivisibility of demand (7) and supply (8) realization.

The general BCBT-I model (1)–(8) can be considered in a few versions, depending on the indivisibility requirements, that may appear only on supply or demand side. For example, if we discard (7) and set $M_d = 1$, for each $d \in D$, then obtained variant of BCBT-I model considers bandwidth indivisibility only from supply point of view, allowing for fully divisible demands realization. Of course, a symmetrical variant can be created by removing constr. (8) and assigning $M_e = 1$, for each $e \in E$. One can notice then, that BCBT-I is an extension of BCBT model, as it would result in the same statement if constrains (7) and (8) would be discarded and all parameters M_e and M_d would be set to 1.

2.2. Main Features of the Model

The BCBT-I is an effective model of bandwidth exchange. Effectiveness is here conceived in the sense of maximizing global economic surplus (market surplus). It is achieved by joint optimization of all submitted buy and sell offers. For given offers the BCBT-I chooses the best allocation of bandwidth determining volumes of accepted offers.

For profit maximizing market players, very important individual goals are the values of economic profits (surplus) they could get. Moreover, from individual player points of view, an important feature of the exchange is the "transparency" and fairness conditions of clearing, which encourage players to place sincere offers and to use truthful bidding strategies reflecting their underlying values. The basic linear BCBT model provides transparent and fair conditions of clearing, since the dual prices in the optimal solution enables setting the competitive market prices for all bandwidths resources on individual links.

In the case of BCBT-I model, the optimal solution determines the realizations of offers that provides efficiency by maximizing the global surplus. However, the MILP (mixed integer linear programming) optimization problem BCBT-I does not provide prices to distribute the surplus between market participants. Thus, a special pricing mechanism for fair economic surplus distribution should be provided. For this purpose the multicommodity balancing mechanism is applied. It is presented in the section below.

3. Multicommodity Balancing Mechanism

A good market mechanism should fulfill many different requirements. From the viewpoint of individual market player several desired properties can be claimed: maximization of individual outcome, individual rationality, impartiality, fairness and simplicity in available strategies [11]. Also from the global point of view, some features of market mechanism are strongly preferable: maximization of social surplus, enabling high competitiveness, budget-balancing, limiting market power, preventing entry deterrence and predation, incentive compatibility. Meeting all these requirements is impossible, what was already proofed in the field of mechanism design theory (Myerson-Satterthwaite impossibility theorem [12]).

In [13] a novel *generic* approach for clearing and fair social welfare distribution in general multicommodity auctions with indivisibility and non-convexities was developed. The method is based on considering two vectors of competitive market clearing sell prices and buy prices of commodities and services. The sell and buy prices are differentiated to share the (non-negative) costs of necessary compensations paid to unfairly priced participants. Sharing of the compensation costs allows the market operator to treat all market participants without discrimination and is justified by incentives that should be given to market participants to bid fairly. The aim is to offset the financial losses and to provide profit optimality to market participants so that they break even. The total compensation cost is calculated in addition to the profit (social welfare) objective function.

The balancing mechanism consists of two steps: allocation and pricing. In the first step a quantity balancing model of the multicommodity auction is solved – the BCBT-I model in our case. In the second step of a balancing mechanism, in order to obtain the best sell and buy competitive prices, the compensation cost is minimized by solving a payment problem.

3.1. Allocation and Payment Rule

Both phases of multicommodity balancing mechanism are performed consecutively. The quantitative balancing is done first. In terms of auction theory it can be treated as a provider of allocation rule that determines an optimal selection of sell and buy bids to realization. The optimal solution maximizes the social welfare and assures zero global profit-optimality loss.

The price determination model can be applied as a separate pricing step of the market clearing procedure, after allocating the resources. It allows the market operator to fairly redistribute the social welfare among market participants, by computing the best buy and sell prices that minimize the costs of necessary (non-negative) compensations. As it was mentioned before, these compensations should be paid to some market participants to avoid individual profit-optimality losses, that may occur due to non-convexities existing in the market.

Problem of setting sell and buy market prices can be formulated as a linear programming task which can be found in [13]. In this article we only present the basic concept of this mechanism.

3.2. Cost of Compensations

There are conflicts between the centralized maximum welfare goal and the profit-driven goals of independent market participants. In particular, the centrally imposed allocation may require some costly offers to be accepted, while rejecting other competitive offers, even though these would have make profit selling the bandwidth under market prices. The rationality assumption of self-interested market participants under competition is that neither buyer or seller can willingly accept a loss of profit, if such loss can be avoided, for example, by reducing consumption, or by making some links unavailable.

On some markets there may exist constraints that significantly limit the trade. In such a case rejecting all non-competitive offers and accepting all competitive offers may lead to suboptimal solutions according to global welfare criterion. In order to maximize social surplus there may be reasonable to impose acceptance (rejection) of some offers that would be rejected (accepted) in the case of market without constraints.

In the analyzed mechanism, there may happen that a market participant is forced by the market operator to buy (sell)

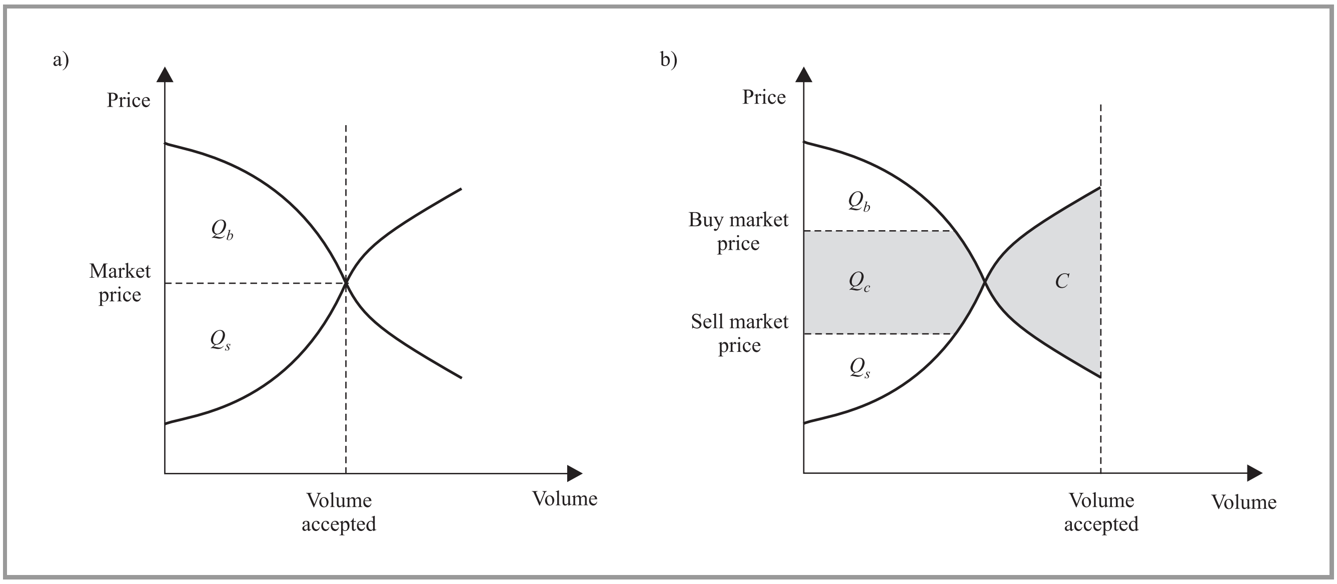


Fig. 1. Supply and demand charts in two cases: equality of sell and buy market price – zero system cost (a); differentiation of sell and buy market prices – non-zero system cost (b).

bandwidth when the market price is higher (lower) than its offered price. To offset the financial losses under market conditions, and to provide the profit optimality to assure break even, the market participant gets the compensation equal exactly to the deficit he or she would have under the market conditions.

Analogously, if a buyer (seller) is competitive with offered price greater (lower) than the market price, but, due to the market operator decisions, the allocated amount of bandwidth is less then expected, then the market participant should get compensation equal to the lost opportunity of gaining surplus.

The overmentioned (non-negative) compensations paid to some market participants to avoid individual profit-optimality losses are minimized in the optimization model, that determines also differentiated market sell and buy price. The prices are differentiated to cover the costs of compensations and to assure budget-balanced property of market mechanism.

3.3. Differentiation of Buy and Sell Market Prices

Different values of sell and buy market price determine which part of social surplus is dedicated for covering the compensations system cost. If system cost of fair global market distribution is zero, then sell and buy market price are the same, corresponding to uniform price at perfect market without any constraints. However, if system cost is non-zero, then buy market price should be greater than sell market price in order to fix appropriate part of social welfare for covering this cost.

In Fig. 1 the supply and demand chart of single commodity is presented. Two different situations are considered: zero system cost and non-zero system cost. In the first case whole social surplus is divided between buyers (Q_b) and sellers (Q_s). There is no need for differentiation of sell and buy market prices, so only one uniform price is set. In the second case system cost of compensations C is not equal zero. Therefore global welfare should be divided for three parts: buyers surplus (Q_b), sellers surplus (Q_s) and surplus dedicated for covering system cost (Q_c). In order to balance the budget, condition $Q_c = C$ must be meet. Determining appropriate size of Q_c can be achieved by differentiation of sell and buy market price of each commodity.

4. Case Study

In this section we present simple examples illustrating the main features of BCBT-I model. Let us consider network with four nodes and four links – Fig. 2. For each link depicted by solid arrow there is one sell offer. Parameters of sell offers are given in the square bracket. First number

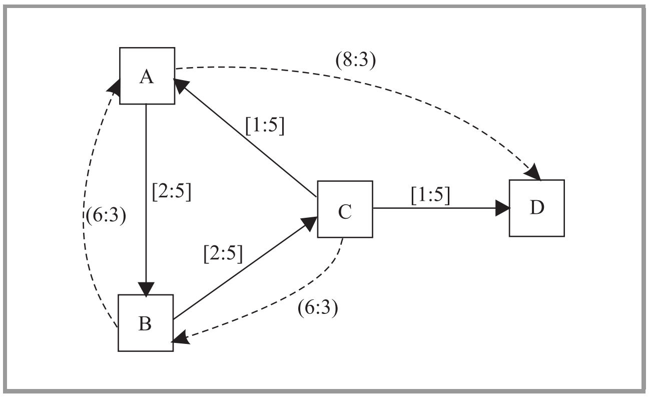


Fig. 2. Network resources with sell and buy offers.

is an unit price and the second number is a maximum capacity. For example, sell offer with unit price 2 and maximum volume 5 is submitted for link connecting nodes A and B. There are also three buy offers, each connected with path depicted by dotted arrows. Parameters of buy offers are given in the parenthesis. First number is an offered price and the second number is a maximum capacity. For example, buy offer with unit price 8 and maximum capacity 3 is submitted for path connecting nodes A and D.

Proposed bandwidth exchange model will be considered in three variants, depending on the bandwidth indivisibility requirements. In order to solve particular variant we use the BCBT and BCBT-I, accordingly.

4.1. Fully Divisible Bandwidth

This variant considers bandwidth as fully divisible commodity. For market balancing, the BCBT model is used. Obtained solution is presented in Fig. 3. For each sell offer the unit price at which seller has sold the bandwidth and accepted volume are given in the brackets – first number denotes price and the second accepted volume. For each buy offer the unit price buyer has to pay and accepted volume are given in parenthesis – first number denotes price and the second accepted volume.

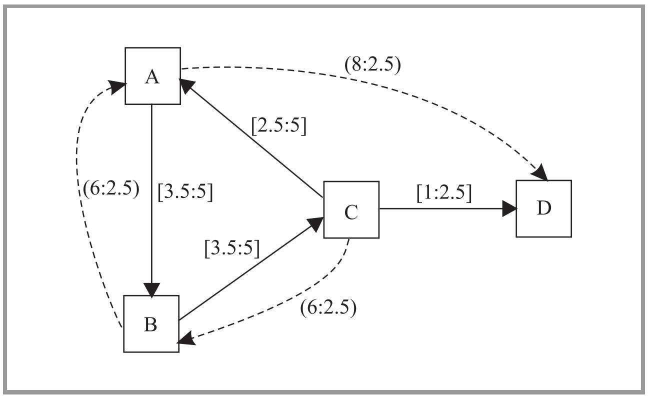


Fig. 3. Solution obtained by the BCBT model in the case of fully divisible bandwidth.

Global welfare obtained in this variant equals 22.5. Its division among market players is fair. Offer at link C-D is partially accepted, therefore price of that link equals market price. Market prices of other links are higher than appropriate offered prices because offers connected with them are fully accepted. All buy offers are partially accepted. Thus, prices that buyers have to pay equal their offered prices.

4.2. Indivisible Bandwidth

This variant considers bandwidth as commodity comprised of several indivisible units. It is assumed that all buy and sell offers are realized in the multiple of unit of size 1, hence all parameters M_e and M_d are set to the value of 1. The solution is presented in Fig. 4. Notation of results is the same as in Fig. 3. Differences between this solution and solution given in previous variant are marked by the bold fold.

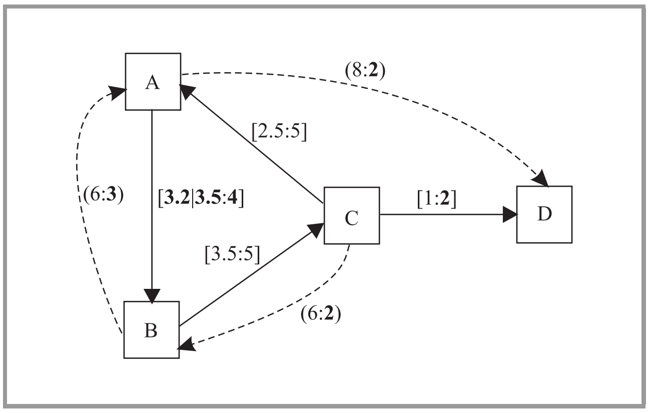


Fig. 4. Solution obtained by the BCBT-I model in the case of indivisible bandwidth.

Global welfare obtained in this variant equals 21. It is lower than in the previous variant, due to bandwidth indivisibility requirement. Market price of link A-B is differentiated: sell market price equals 3.2 and buy market price equals 3.5. Because volume of offer submitted for that link equals 4, the part of social welfare dedicated for covering system cost equals $4 \cdot (3.5 - 3.2) = 1.2$. It is used to pay compensation to owner of link A-B. Note that although the offered sell price (2) is lower than market sell price of that link (3.2), the sell offer is not fully accepted. The maximum volume is 5 while accepted volume is 4. Owner of this offer faces profit opportunity loss equals $(5 - 4) \cdot (3.2 - 2) = 1.2$. The multicommodity balancing mechanism results in maximal global welfare achieved with relatively small system cost.

4.3. Indivisible Bandwidth on the Supply Side Only

This variant assumes that the constraint of bandwidth indivisibility is required only on the supply side. All sell offers are realized in the multiple of unit of size 1, hence all parameters M_e equal 1. Modified variant of BCBT-I, without constraints (7) and with all parameters M_d set to 1, gives the solution presented in Fig. 5. Notation of results is the same as in Fig. 3. Differences between this solution and solution given by the BCBT model are marked by the bold fold.

Global welfare obtained in this variant equals 22. It is lower than in the first variant considering fully divisible bandwidth and greater than in the second variant considering indivisible bandwidth on both supply and demand side. When we compare the allocations, we can see that the accepted volume at link C-D equals 2.5 in case of the BCBT model and 3 in this case. All others values of volumes are the same in both cases. As total demand equals total supply in the case of the BCBT model, then in this case supply of bandwidth is higher than bandwidth demand. The cost of the excessive supply equals $(3-2.5)\cdot 1 = 0.5$.

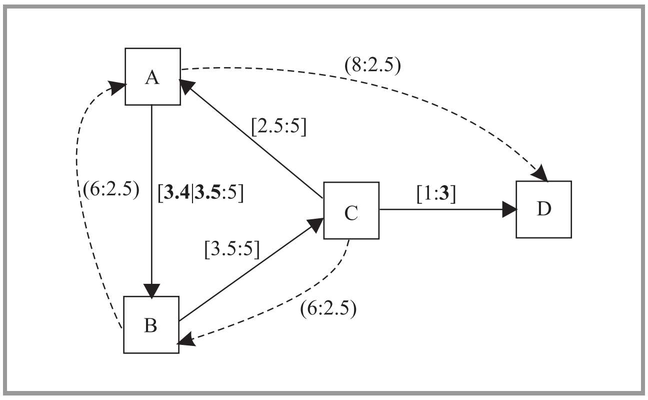


Fig. 5. Solution obtained by variant of the BCBT-I model assuming indivisible bandwidth on the supply side only.

When we compare market price of link A-B we can see that in the first variant price equals 3.5. In this variant market price of that link is differentiated: sell market price equals 3.4 and buy market price equals 3.5. Although there are no system cost of fair global welfare division, the differentiation of buy an sell market price allows to cover cost $5\cdot(3.5-3.4) = 0.5$ of superfluous supply.

5. Summary

In this paper we presented the multicommodity auction model BCBT-I for indivisible network resource allocation. It is an extension of the BTBT market model that considers bandwidth as fully divisible commodity. Constraints assuring bandwidth indivisibility causes that the BCBT-I model is formulated as a mixed-integer problem. The BCBT-I model ensures determining optimal (according to global welfare) volume of accepted offers. For fair distribution of social surplus additional mechanism, we examined the balancing mechanism based on differentiation of sell and buy market price. Illustrative examples confirm the proposed model accuracy. Further works are needed to analyze the computational complexity of the proposed model. The pricing issues introduced in this paper needs also further, comprehensive studies.

Acknowledgments

The authors acknowledge the Ministry of Science and Higher Education of Poland for partially supporting the research through project PBZ-MNiSW-02/II/2007.

References

[1] G. Cheliotis, “Structure and dynamics of bandwidth markets”, Ph.D. thesis, N.T.U., Athens, 2001.

[2] S. Bessler and P. Reichl, *A Network Provisioning Scheme Based on Decentralized Bandwidth Auctions*. Operations Research/Computer Science Interfaces Series. New York: Springer, 2006.

[3] R. Rabbat and T. Hamada, “Revisiting bandwidth-on-demand enablers and challengers of a bandwidth market” in *Netw. Oper. Manage. Symp. NOMS 2006*, Vancouver, Canada, 2006, pp. 1–12.

[4] C. Courcoubetis, M. P. Dramitinos, and G. D Stamoulis, “An auction mechanism for bandwidth allocation over paths”, in *Int. Teletraf. Congr. ITC-17*, Salvador da Bahia, Brazil, 2001, pp. 1163–1174.

[5] M. Dramitinos, G. D. Stamoulis, and C. Courcoubetis, “An auction mechanism for allocating the bandwidth of networks to their users”, *Comput. Netw.*, vol. 51, pp. 4979–4996, 2007.

[6] A. Lazar and N. Semret, “Design and analysis of the progressive second price auction for network bandwidth sharing”, *Telecommun. Syst.* (Special Issue on Network Economics), 1999.

[7] Ch. Kaskiris, R. Jain, R. Rajagopa, and P. Varaiya, “Combinatorial auction bandwidth trading: an experimental study”, in *Developments on Experimental Economics. Lecture Notes in Economics and Mathematical Systems*, vol. 590. Berlin: Springer, 2007, pp. 181–186.

[8] W. Stańczuk, J. Lubacz, and E. Toczyłowski, “Trading links and paths on a communication bandwidth markets”, *J. Univer. Comput. Sci.*, vol. 14, no. 5, pp. 642–652, 2008.

[9] R. Jain and P. Varaiya, “Combinatorial bandwidth exchange: mechanism design and analysis”, *Commun. Inform. Sci.*, vol. 4, no. 3, pp. 305–324, 2004.

[10] P. Pałka, K. Kołtyś, E. Toczyłowski, and I. Żółtowska, “Model for balancing aggregated communication bandwidth resources”, in *7th Int. Conf. Decis. Supp. Telecommun. Inform. Soc. DSTIS 2008*, Warsaw, Poland, 2008 (*J. Telecommun. Inform. Technol.*, 2009 – to appear).

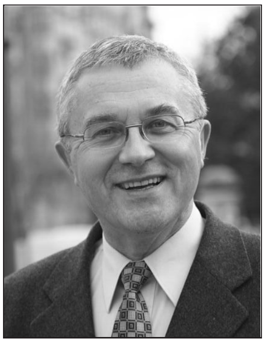
[11] P. Klemperer, *Auctions: Theory and Practice*. Princeton: Princeton University Press, 2004.

[12] R. B. Myerson and M. A. Satterthwaite, “Efficient mechanisms for bilateral trading”, *J. Econom. Theory*, vol. 28, pp. 265–281, 1983.

[13] E. Toczyłowski, *Optymalizacja procesów rynkowych przy ograniczeniach*. Warsaw: EXIT, 2003 (in Polish).

Kamil Kołtyś received the M.Sc. degree in computer science in 2007 from the Warsaw University of Technology, Poland. Currently he prepares his Ph.D. thesis in computer science at the Institute of Control and Computation Engineering at the Warsaw University of Technology. His research interests include decision support, optimization and bandwidth trading. His current research is focused on application of multicommodity turnover models to network resource allocation.
e-mail: K.J.Koltys@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Piotr Pałka received the M.Sc. degree in 2005 from the Warsaw University of Technology, Poland. Currently he prepares his Ph.D. thesis in computer science at the Institute of Control and Computation Engineering at the Warsaw University of Technology. His research interest is incentive compatibility on the infrastructure markets. His current research is focused on application of multicommodity turnover models.
e-mail: P.Palka@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Eugeniusz Toczyłowski is Professor, the Head of Operations Research and Management Systems Division, at the Institute of Control and Computation Engineering at the Warsaw University of Technology, Poland. He received the M.Sc. degree in 1973, Ph.D. in 1976, D.Sc. in 1989, and the title of Full Professor in 2004. His main research interests are centered around the operations research models and methods, including structural approaches to large scale and discrete optimization, auction theory and competitive market design under constraints, multicommodity trading models, and design of management information systems.
e-mail: E.Toczylowski@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Izabela Żółtowska received the M.Sc. degree in 2000 and the Ph.D. degree in 2006 from the Warsaw University of Technology, Poland. She is an Assistant Professor at the Institute of Control and Computation Engineering at the Warsaw University of Technology. Her research focuses on the optimization models applied on the restructured competitive markets.
e-mail: I.Zoltowska@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Spectral Division of the Optical Fiber Passband Using Narrowband Controllable Filter on the Base of Semiconductor Waveguide Microresonator

Igor Goncharenko, Alexander Esman, Grigory Zykov, Vladimir Kuleshov, Marian Marciniak, and Vladimir Pilipovich

Abstract—We analyze the new principle of multichannel spectral division of optical fiber passband using controllable narrowband integrated optical filters composed of two-coupled ring microresonators made of different semiconductor materials. It is shown that appropriate selecting the semiconductor material and optimizing the design factors of selective optical element allows creating the simple and economical integrated optical filter with bandwidth 0.1 nm, frequency separation between adjacent optical carriers 0.2 nm and signal-to-noise ratio 50 dB. Utilizing such filters in optical fiber communication lines makes it possible to increase the number of transmitted in parallel optical carrier wavelengths up to 160 and even more, i.e., to provide the traffic transmission with the speed up to 1.6 Tbit/s in one direction and in single optical fiber.

Keywords—carrier injection, controllable optical filter, coupled waveguides, optical passband, resonance wavelength, resonator optical length, ring microresonator.

1. Introduction

The rapid growth of the needs of modern society in large data streams, and first of all the Internet, stimulates the research oriented on increasing the carrying capacity of optical communication channels. Wavelength division multiplexing (WDM) and dense wavelength division multiplexing (DWDM) technologies are increasingly effectively used for construction of high-speed main lines and optical communication networks. The permanent extension of carrying capacity of optical fiber communication lines on that way arises due to application of the latest achievements of theoretical and experimental research from the one hand, and new achievements of optical technology time division multiplexing (TDM) from the other hand [1–4].

The parallel transmission of N data streams on corresponding carrier optical wavelength $\lambda_1 \ldots \lambda_N$ allows extending the carrying capacity of optical communication lines based on WDM/DWDM technologies by adding progressively the new optical channels as the network develops. The narrowband optical filter with the passband controllable with the high speed is the key element of such devices and is used for spectral multiplexing/demultiplexing of optical channels.

In present paper we consider the multichannel spectral division of optical fiber passband on the base of microresonators made from different semiconductor materials. The physical essence of the method is in shifting the microresonator resonance wavelength because of the changing its optical length by varying the free charge carriers density influencing on the material refractive indices [5].

2. Structure of the Filter and Method of Calculation of its Parameters

The structural diagram of the proposed narrowband controllable integrated optical filter is shown in Fig. 1. The filter constitutes two sequentially optically coupled ring waveguide microresonators with band radius of tens of microns, which disposed on the distance 200 nm from each other and from straight input and output optical waveguides. The interaction length and gap width between the waveguides define their optical coupling coefficient. The narrow optical frequency bands corresponding to the resonance frequencies couple from the input waveguide to the microresonator [6, 7]. By changing the resonance conditions (for instance, by changing the resonator optical length, i.e., its geometrical length or waveguide effective index) one can vary the frequency band coupled into resonator. The effective index of the waveguide made from semiconductor materials can be changed by optical or electrical injection of the free carriers [5, 8]. For electrical injection the n-doped regions are created outside the ring microresonators while the p-doped regions are disposed inside the rings. When the electrical voltage is supplied on such diode structure the electrons and holes penetrate into the waveguide material and change its effective index and thereby the resonance frequency.

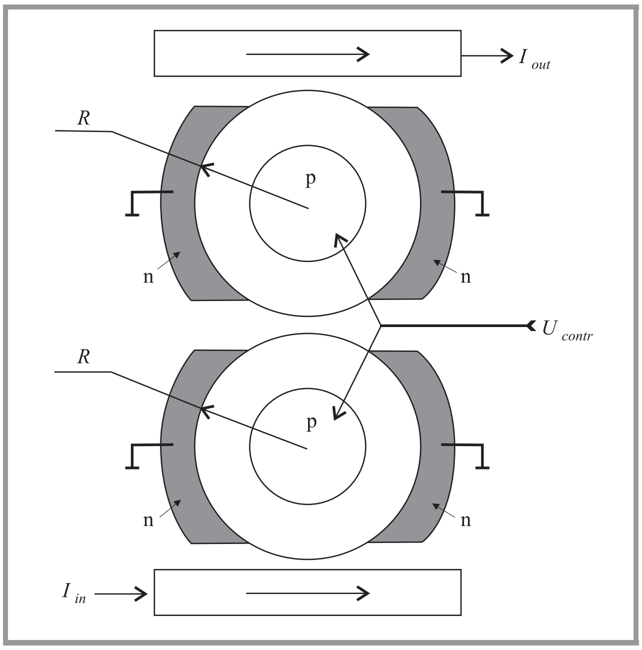


Fig. 1. The structural diagram of controllable integrated optical filter on the base of two optically coupled waveguide ring microresonators.

We implement the numerical simulations for analysis of the parameters of controllable integrated optical filter on the base of ring waveguide microresonators. In order to obtain the electromagnetic fields distribution on the filter input and output corresponding to different time points, resonance and transfer characteristics of the resonator we solve the wave equation written for Borgnis' electrical function [9]. That allows us to content only the considering E_z component of transverse magnetic-wave for the structure composed of two straight waveguides and optically sequentially coupled ring microresonators and in this way to apply the d'Alembert's equation in Cartesian coordinates [10]

$$\frac{\partial^2 E_z}{\partial x^2}+\frac{\partial^2 E_z}{\partial y^2}-\frac{n_S^2}{c^2}\frac{\partial^2 E_z}{\partial t^2}=0 \tag{1}$$

for straight waveguides and in cylindrical coordinates:

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial E_z}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2 E_z}{\partial\varphi^2}-\frac{n_R^2}{c^2}\frac{\partial^2 E_z}{\partial t^2}=0 \tag{2}$$

for ring microresonators, where n_S and n_R are the effective refractive indices of the input and output waveguides and ring microresonators, respectively.

The input and boundary conditions complete the equations (1, 2). For solving the wave equations (1, 2) with three variables we use the explicit numerical model of the "cross" type [11]. In accordance with this model the spatial and time derivatives of the second order in wave equation in Cartesian and cylindrical coordinates Eqs. (1, 2) are substituted for

$$\frac{\partial^2 E_z}{\partial x^2}\approx\frac{E_z(x_{i+1},y_j,t_n)-2E_z(x_i,y_j,t_n)+E_z(x_{i-1},y_j,t_n)}{(\Delta x)^2}, \tag{3}$$

$$\frac{\partial^2 E_z}{\partial y^2}\approx\frac{E_z(x_i,y_{j+1},t_n)-2E_z(x_i,y_j,t_n)+E_z(x_i,y_{j-1},t_n)}{(\Delta x)^2}, \tag{4}$$

$$\frac{\partial^2 E_z}{\partial t^2}\approx\frac{E_z(x_i,y_j,t_{n+1})-2E_z(x_i,y_j,t_n)+E_z(x_i,y_j,t_{n-1})}{(\Delta t)^2}, \tag{5}$$

$$\frac{\partial E_z}{\partial\rho}\approx\frac{E_z(\rho_{i+1},\varphi_j,t_n)-E_z(\rho_i,\varphi_j,t_n)}{\Delta\rho}, \tag{6}$$

$$\frac{\partial^2 E_z}{\partial\rho^2}\approx\frac{E_z(\rho_{i+1},\varphi_j,t_n)-2E_z(\rho_i,\varphi_j,t_n)+E_z(\rho_{i-1},\varphi_j,t_n)}{(\Delta\rho)^2}, \tag{7}$$

$$\frac{\partial^2 E_z}{\partial\varphi^2}\approx\frac{E_z(\rho_i,\varphi_{j+1},t_n)-2E_z(\rho_i,\varphi_j,t_n)+E_z(\rho_i,\varphi_{j-1},t_n)}{(\Delta\varphi)^2}, \tag{8}$$

where Δx, Δy, $\Delta\rho$ and $\Delta\varphi$ are the spatial differences, Δt is time difference, i, j and n are integer numbers.

The input and boundary conditions complete the equations (1, 2). The calculations are carried out for definite region, thus the wave function on the region boundaries is set equal to zero.

The signal on the waveguide input is set as

$$E_z(x_0=0,y,t)=E_0\exp\left(-\frac{(y-y_0)^2}{a^2}\right)\sin(2\pi f t), \tag{9}$$

and the distribution of the E_z component of the electromagnetic field in interaction region of the input waveguides and ring microresonators is calculated as

$$E_z(\rho,\varphi,t)=E_1\exp\left(-\frac{\big((\rho-\rho_0)\cos\varphi\big)^2}{a^2}\right)\sin(2\pi f t), \tag{10}$$

where E_0 is the amplitude of the input signal in the first waveguide, E_1 is amplitude of the signal passing into the ring microresonator from input waveguide, f is carrier frequency, values x_0, y_0, ρ_0, φ_0 ($\varphi_0=0$) and a define the shape and spatial position of input Gaussian functions.

The spatial intervals Δx and Δy on coordinate axes is set less than the input radiation wavelength, and for the choice of discretization time step the Courant' generalized stability condition [11]

$$\Delta t\le\frac{1}{c\sqrt{(1/\Delta x)^2+(1/\Delta y)^2}} \tag{11}$$

is taken into account. The relation of variables for solving the wave Eq. (2) in cylindrical coordinates is set the similar way.

The coupling coefficient between input/output waveguides and ring microresonators is k_S and the one between two ring waveguides is k_R.

The resonance characteristic of the structure under consideration is defined as

$$T(\lambda) = \frac{I_{out}}{I_{in}} = \frac{E^2_{z\ out}}{E^2_{z\ in}}, \tag{12}$$

where $E_{z\ in}$ and $E_{z\ out}$ are the amplitudes and I_{in} and I_{out} are the intensities of the filter input and output signals, respectively.

In numerical modeling we use the next waveguide parameters: the length of straight input and output waveguides is 13 μm, the waveguides thickness and width are 0.3 μm, separation of input/output waveguides and ring microresonators is 0.2 μm.

3. Results and Discussion

We have applied the algorithm described above to analyze the switching and resonance characteristics of the filter proposed. The results are plotted in Figs. 2–4.

Figure 2 shows the dependence of the FWHM (full width at half maximum) $\Delta\lambda$ of single resonance line on ring radius for filter with one (Fig. 2a) and two (Fig. 2b) ring microresonators calculated for three semiconductor materials: Si ($n = 3.483$, curves *1*), GaAs ($n = 3.2$, curves *2*) and InP ($n = 3.172$, curves *3*).

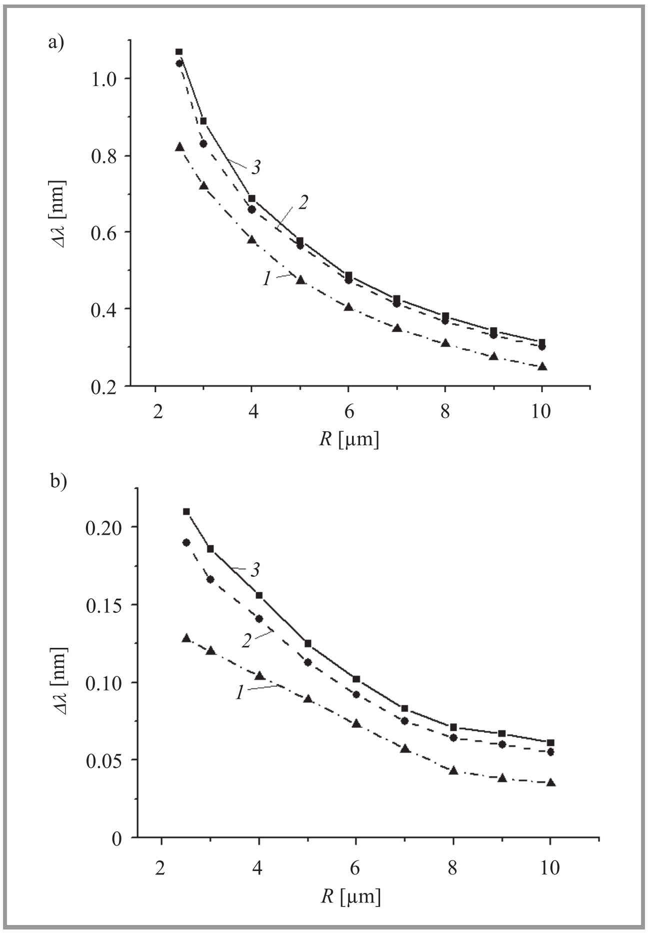


Fig. 2. Dependence of the FWHM $\Delta\lambda$ of the filter with one (a) and two (b) ring microresonators made from Si (*1*), GaAs (*2*) and InP (*3*) on ring radius *R*.

The passband FWHM of the filter made from GaAs on the base of one ring microresonator with radius $R = 2.5$ μm is 1.04 nm (curve *1* in Fig. 2). This result is in a good agreement with the one reported in [10]. The passband width of the same filter on the level of 0.1 of the maximum is 1.72 nm. Thus for effective switching the resonance band it is necessary to shift the wavelength of its maximum on the spectral interval of the order of 1.7 nm. Our calculations show that for such shifting the passband of the filter with ring radius $R = 2.5$ μm it is necessary to change the waveguide material index in second digit after comma, that is practically unrealizable [5, 8]. In practice, in controllable integrated optical filters from GaAs the real changing the free carrier density can be as high as $2.5 \cdot 10^{18}$ cm^{-3} [5]. This results in variation of waveguide material index on the value up to $\Delta n = 0.003$. For filter with ring radius $R = 10$ μm the passband width on the level of 0.1 of the maximum is 0.48 nm. That means that in order to shift the resonance passband on the value comparable with its width the waveguide effective index has to be changed on the value $\Delta n \approx 0.002$, which can be realize in practice [5, 8]. The similar conclusion is valid for the filter made from Si and InP, in which the change of free carrier density on the value $5 \cdot 10^{18}$ cm^{-3} [12] and $3 \cdot 10^{18}$ cm^{-3} [13], respectively, leads to index variation Δn approximately equal 0.004 and 0.003 [13, 14].

The passband control efficiency of the integrated optical filter based on one and two optically coupled microresonators is estimated by the value of the ratio η of maximal intensities of its output signals in two positions: $\eta = I_{on}/I_{off}$, where I_{on} and I_{off} are the maximal intensities corresponding the open and closed filter conditions, respectively. Our calculations show that the filters with one resonator with the radius in the range $R = 2.5 \ldots 10$ μm can't be used in the most of practical applications because of the small value of η.

Figure 3 shows the dependence of the ratio η on Δn value for the filters composed of single-ring resonator with $R = 14$ μm (Fig. 3a) and two-ring resonator with $R = 10$ μm (Fig. 3b) made from Si ($n = 3.483$, curves *1*), GaAs ($n = 3.2$, curves *2*) and InP ($n = 3.172$, curves *3*). The single-ring microresonator and two-rings filter of such size occupies the same substrate area.

The modeling of the spectral characteristic of the filter composed of two-coupled microresonators on the base of GaAs and Si shows that their resonance passband width is narrower in more than 4 times as compared with single-ring filter made from the same materials and with the same radii (see Fig. 2).

Figure 4 shows the calculated resonance passbands of the filters composed of two similar optically coupled microresonators from Si (Fig. 4a), GaAs (Fig. 4b) and InP (Fig. 4c) in initial condition (curves *1*) and in switched condition ($\Delta n = 0.004$, curves *2*). The dependence of the ratio η on the variation of the material indices Δn for cases of Fig. 4 is presented on Fig. 3b by curves *1*, *2* and *3*, respectively. For the filters under consideration the high value η exceeding 10^5 is achieved for Δn equal $4 \cdot 10^{-4}$ (curve *3*) for $R = 10$ μm.

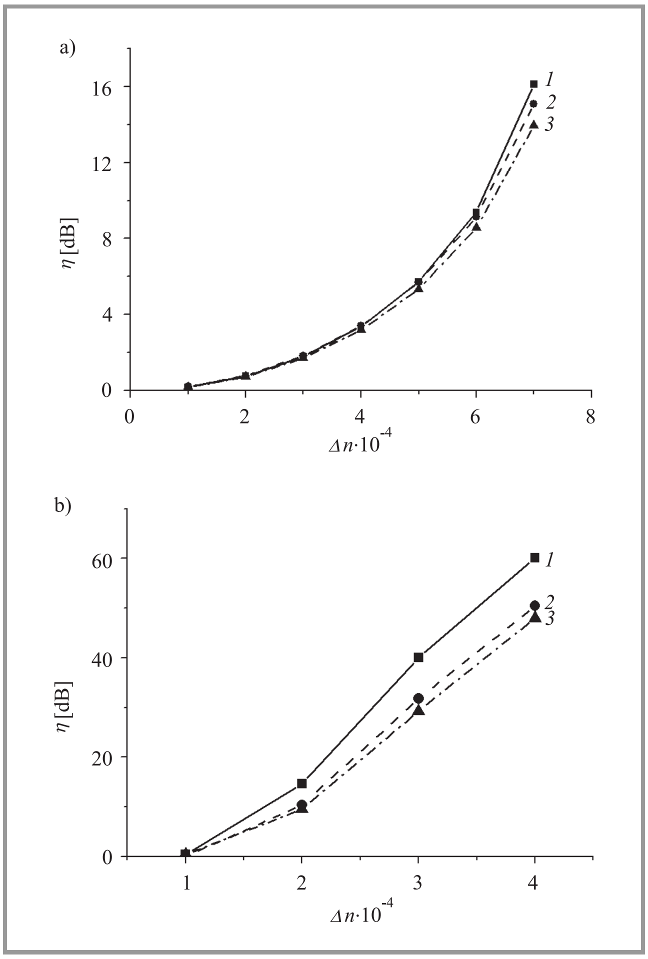


Fig. 3. Dependence of the ratio η on Δn value for the filter with one ring microresonator with $R = 14$ μm (a) and two (b) ring microresonators with $R = 10$ μm made from Si (*1*), GaAs (*2*) and InP (*3*).

The speed of response of the filter under consideration is defined by the sum of the transition time of the output signal and relaxation (recombination) time of the charge carriers. The signal establishing in single-ring microresonator is accomplished in approximately 29 passing the radiation through the microresonator ring with radius $R = 10$ μm and coupling coefficients between straight and ring waveguides equal 0.5 [7]. That amounts approximately 20.4 ps for GaAs, 18.8 ps for Si and 18.6 ps for InP. The output signal establishing in the filter from two coupled microresonators is accomplished in 26.3, 24.1 and 23.9 ps for GaAs, Si and InP, respectively.

The relaxation (recombination) time of charge carriers in Si, GaAs and InP is 23 ps [15], 0.4 ps [16] and 0.2 ps [17], respectively. Thus the maximal signal-repetition frequency for optical filter on the base of one microresonator ring is equal to 12.0 GHz for silicon waveguide, 24.0 GHz for gallium arsenide waveguide and 26.6 GHz for indium phosphide waveguide. For the filter composed of two optically coupled rings with radius 10 μm the maximal signal-repetition rates are smaller: 10.8, 19.3 and 21.3 GHz for Si, GaAs and InP, respectively.

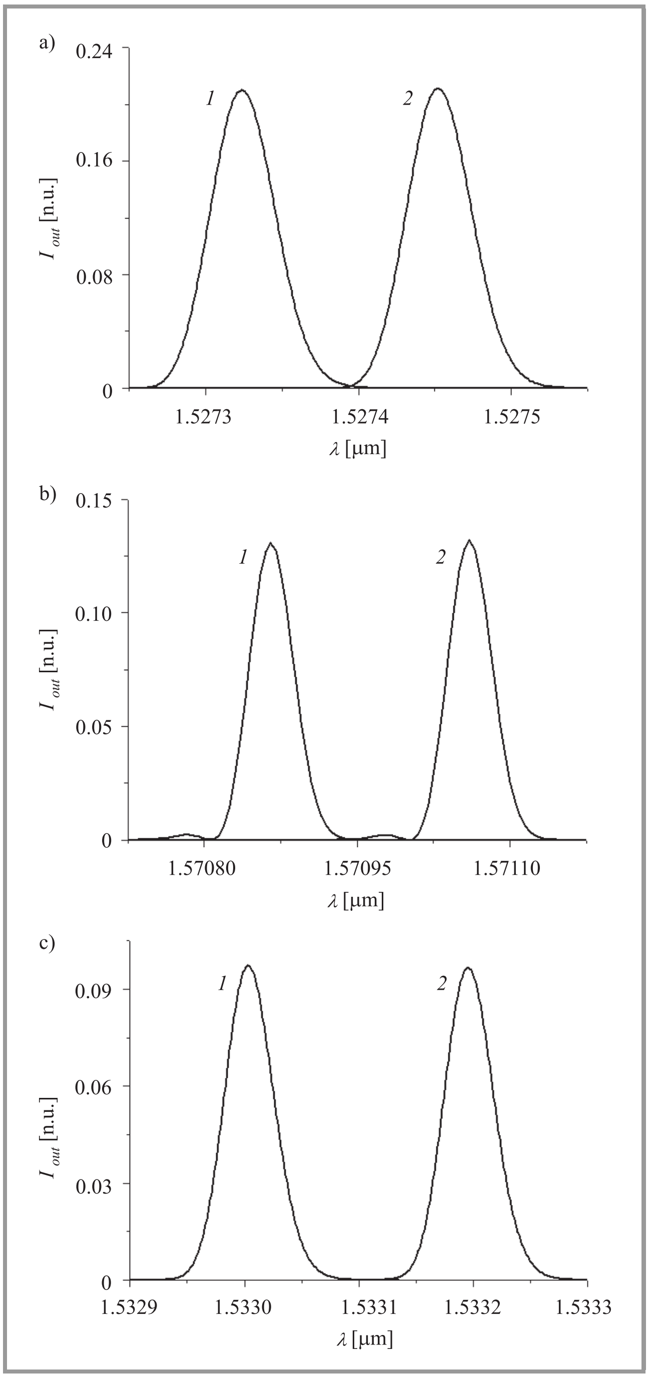


Fig. 4. Resonance passbands of the filters composed from two similar optically coupled microresonators with $R = 10$ μm made from Si (a) for $n = 3.483$ (*1*) and $n = 3.4833$ (*2*); GaAs (b) for $n = 3.2$ (*1*) and $n = 3.2004$ (*2*); and InP (c) for $n = 3.172$ (*1*) and $n = 3.1724$ (*2*).

4. Conclusions

We propose the numerical modeling the controllable optical filter composed of two optically coupled semiconductor waveguide ring microresonators disposed between two straight input/output waveguides. We compare the passband width and shift of maximums of resonance bands of the filters made of different semiconductor materials in dependence on variation of their refractive indices. It is shown that the signal-to-noise ratio on the output of the filter proposed could be as high as 50 dB for index variation 0.004, which can be realized in practice.

Speed of response of the filter made from GaAs or InP is about 20 GHz and in two times larger that the pulse-repetition rate of silicon filter (about 10 GHz). However, Si is relative cheap material and its technology is well developed with low rejection rate. Therefore, gallium arsenide or indium phosphide filters could be used for high-performance information processing while in mass production optoelectronic interfaces one can apply the silicon filters.

The filter passband width on the level of 0.1 of maximum is of the order of 0.1 nm. The bandwidth of element base of modern optical fiber communication lines is 32 nm. Thus the use of such filters allows realizing the parallel data transmission over the 160 channels in single fiber. When the speed of response of single channel is 10 Gbit/s the communication line capacity could achieve 1.6 Tbit/s.

The proposed optical filter can contribute to advancing the optical fiber communication lines using WDM and DWDM technologies and operating on multigigabit and terabit velocities.

Acknowledgements

Alexander Esman, Vladimir Kuleshov and Vladimir Pilipovich thank the Belarussian Fund of Fundamental Research (grant F07MS-004) for the financial support of this work. The authors thank the European Project COST Action MP0702 "Towards functional sub-wavelength photonic structures" for stimulating interactions.

References

[1] I. A. Mamzelev, V. M. Malafeev, A. D. Snegov, and L. V. Yurasova, *Technologies and Equipment*. Moscow: Eko-Trends, 2005.

[2] V. E. Kuznetsov *et al*., "Method of redundancy in phase-locked optical communication line with system of spectral multiplex". Patent of Russian Federation, no. 2307469.

[3] V. G. Tatsenko and A. K. Shishov, "Systems with channel spectral multiplexing (WDM and DWDM systems) in optical fiber systems for communication of information". Part 2, *Tele-Sputnik*, no. 2, pp. 24–29, 2004.

[4] A. V. Shmal'ko, "Systems for spectral multiplex of optical channels", *Bull. Commun.*, no. 4, pp. 162–170, 2002.

[5] T. A. Ibrahim *et al*., "Lightwave switching in semiconductor microring devices by free carrier injection", *J. Lightw. Technol.*, vol. 21, no. 12, pp. 2997–3003, 2003.

[6] B. E. Little *et al*., "Ultra-compact Si-SiO microring resonator optical channel dropping filters", *IEEE Photon. Technol. Lett.*, vol. 10, no. 4, pp. 549–551, 1998.

[7] I. A. Goncharenko, A. K. Esman, V. K. Kuleshov, and V. A. Pilipovich, "Optical broadband analog-digital conversion on the base of microring resonator", *Opt. Commun.*, vol. 257, no. 1, pp. 54–61, 2006.

[8] S. Abdalla *et al*., "Carrier injection-based digital optical switch with recon-figurable output waveguide arms", *IEEE Photon. Technol. Lett.*, vol. 16, no. 4, pp. 1038–1040, 2004.

[9] S. T. Chu and S. K. Chaudhuri, "Numerical modeling the electromagnetic field distribution in the waveguide based on photonic crystal", *J. Lightw. Technol.*, vol. 7, pp. 2033–2038, 1989.

[10] A. S. Loginov and A. S. Majorov, "Numerical modeling the characteristics of selective integrated optical elements taking into account the loss compensation", *J. Radioelectron.*, no. 3, pp. 17–21, 2007.

[11] N. N. Kalitkin, *Numerical Methods*. Moscow: Nauka, 1978, pp. 425–439.

[12] M. S. Bressler, O. B. Gusev, and E. I. Terukov, "Silica edge electroluminescent: heterostructure amorphous silicon-crystalline silicon", *Sol. Phys.*, vol. 46, no. 1, pp. 18–20, 2004.

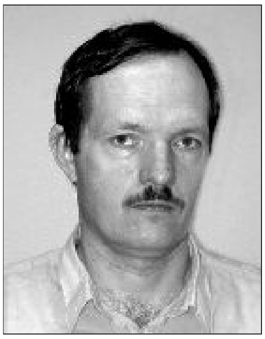
[13] M. V. Kotlyar, L. O'Faolain, A. B. Krysa, and T. F. Krauss, "Electrooptic tuning of InP-base microphotonic Fabry-Perot filters", *J. Lightw. Technol.*, vol. 23, no. 6, pp. 2169–2174, 2005.

[14] C. Manolatou and M. Lipson, "All-optical silicon modulators based on carrier injection by two-photon absorption", *J. Lightw. Technol.*, vol. 24, no. 3, pp. 1433–1439, 2006.

[15] F. Y. Gardes *et al*., "Micrometer size polarization independent depletion-type photonic modulator in silicon on insulator", *Opt. Expr.*, vol. 15, no. 9, pp. 5879–5884, 2007.

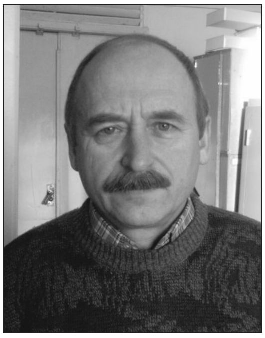
[16] S. S. Strelchenko and V. V. Lebedev, *Compounds A^3B^5*. Handbook. Moscow: Metallurgy, 1984, p. 60.

[17] A. S. Tager, "Prospects of application of indium phosphide in microwave semiconductor electronics", in *Indium Phosphide in Semiconductor Electronics*, S. I. Radautsan, Ed. Kishinev: Stiintza, 1988, pp. 120–132.

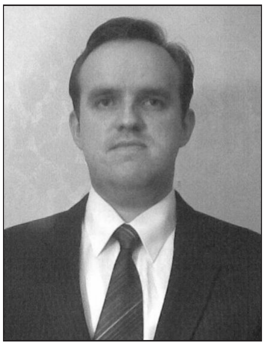


Igor A. Goncharenko was born in Minsk, Belarus. He received M.Sc. degree from Belarussian State University in 1981, Ph.D. degree in physics and mathematics from the USSR Academy of Sciences (Moscow) in 1985 and Dr.Sc. degree from the National Academy of Sciences of Belarus (Minsk) in 2001. Since September 2007 he takes up the position of Professor of the Department of Natural Sciences in the Institute for Command Engineers of the Ministry of Emergencies of the Republic of Belarus and part time Professor position in Belarussian National Technical University. His fields of investigation include theory of complicated optical fibre and waveguide devices, optical information processing.

e-mail: Igor02@tut.by
Institute for Command Engineers
of the Ministry of Emergencies of the Republic of Belarus
Department of Natural Sciences
Mashinostroiteley st 25
220118 Minsk, Belarus



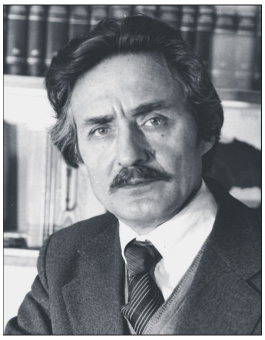
Alexander Konstantinovich Esman was born in 1951. He got his M.Sc. from Belarusian State University in 1973 and Ph.D. in physical and mathematical sciences in 2003. He is a major researcher of the B. I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus. His concept of scientific research is optical data processing.
e-mail: lomoi@inel.bas-net.by
Institute of Physics
National Academy of Sciences
22 Logoisky Trakt
220090 Minsk, Belarus



Grigory Lyutsianovich Zykov was born in 1980. He got his M.Sc. from Gomel State University in 2002. He has been a candidate of technical sciences since 2007. He is a researcher of the B. I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus. His concept of scientific research is optical data processing, simulation of the physical processes in the dielectric and semiconductor materials.
e-mail: lomoi@inel.bas-net.by
Institute of Physics
National Academy of Sciences
22 Logoisky Trakt
220090 Minsk, Belarus

Vladimir Konstantinovich Kuleshov was born in 1951. He got his M.Sc. from Belarusian State University in 1973. He has been a candidate of technical sciences since 1984. He is a leading researcher of the B. I. Stepanov Institute of Physics of the National Academy of Sciences of Belarus. His concept of scientific research is optical data processing and elements of integrated computer engineering.
e-mail: lomoi@inel.bas-net.by
Institute of Physics
National Academy of Sciences
22 Logoisky Trakt
220090 Minsk, Belarus

Marian Marciniak graduated in solid state physics from Marie Curie-Skłodowska University in Lublin, Poland, in 1977. He holds a Ph.D. degree in optoelectronics (1989), and a Doctor of Sciences (Habilitation) degree in physics/optics (1997). Actually he is the Head of Department of Transmission and Optical Technologies at the National Institute of Telecommunications. He authored 280 publications, including a number of invited conference presentations. He serves as a Honorary International Advisor to the George Green Institute for Electromagnetics Research, University of Nottingham, UK. He serves as the Chairman of the Management Committee of COST Action MP0702 "Towards functional sub-wavelength photonic structures".
e-mail: M.Marciniak@itl.waw.pl
e-mail: marian.marciniak@ieee.org
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland



Vladimir Antonovich Pilipovich was born in Gomel region, Belarus, in 1931. He received M.Sc. degree from Belarussian State University in 1954, Ph.D. degree in physics and mathematics in 1959 and Dr.Sc. degree in 1972. He has the rank of Professor in physics and is a Member of the National Academy of Sciences of Belarus (NASB). He has worked in NASB since 1957. From 1973 to 1998 he was a Director of the Institute of Electronics. At present he is a Chief Research Fellow of the Institute of Physics NASB. He is a co-author of more than 300 scientific publications including 3 monographs. Field of his scientific interests includes lasers and laser physics, optoelectronics and optical information processing. He is a Member of the IEEE and SPIE.
e-mail: lomoi@inel.bas-net.by
Institute of Physics
National Academy of Sciences
22 Logoisky Trakt
220090 Minsk, Belarus

Paper

On the Effects of Code Cardinality for TH-PPM Ultra Wideband Systems

Mohammad Upal Mahfuz, Kazi M. Ahmed, and Nandana Rajatheva

Abstract—This paper demonstrates the effects of code cardinality at transmitter section on bit error rate (BER) performance of time hopping pulse position modulation (TH-PPM) based ultra wideband (UWB) indoor radio communication. In the transmitter, different code cardinality values have been chosen and correspondingly the effects on BER of the system have been investigated. The recently accepted IEEE 802.15.3a model of the UWB channel has been used as the propagation channel model in indoor environment. Results show that the system BER performance is significantly dependent on the code cardinality value of time hopping code. For such higher code cardinality values as in the range from 30 to 50, the BER performance degrades severely. Finally, code cardinality in the range from 10 to 15 has been recommended for TH-PPM system in UWB indoor communications providing better BER performance for the same system data rate requirement.

Keywords—code cardinality, multiple-user interference, TH-PPM, ultra wideband.

1. Introduction

Ultra wideband (UWB) transmission systems are seen as very promising solutions for many wireless indoor and short-range communication environments such as ad hoc, home, personal and body area networks, since they can exploit the 3–10 GHz unlicensed spectrum [1]. Impulse radio UWB (IR-UWB) systems are one of the two main streams of UWB technology. In IR-UWB systems, subnanosecond pulses are transmitted in the wireless medium. For time hopping (TH) UWB radio, it is indicated that both modulation method and TH code influence the bit error rate (BER) performance of the system [2]. However, the effect of TH code properties directly on the BER performance still lacks enough research attention. Code cardinality is an important characteristic in case of TH code used for time hopping pulse position modulation (TH-PPM) systems. This is why the effect of code cardinality on BER is a significant investigation in this field. Time hopping is a popular multiple access scheme used especially with pulse position modulation (PPM). For a typical time hopping format employed by an energy normalized impulse radio signal, the output signal of the kth transmitter can be expressed as

$$s_{tr}^{(k)}\left(t^{(k)}\right) = \sum_{j=-\infty}^{\infty} w_{tr}\left(t^{(k)} - jT_f - c_j^{(k)} T_c - \delta d_{\lfloor j/N_s \rfloor}^{(k)}\right), \quad (1)$$

where $c_j^{(k)}$ is the distinct time hopping sequence, t is the transmitter clock time, w_{tr} is the transmitted monocycle waveform, T_f is the pulse repetition time or the frame time, T_c is the chip duration, δ is the modulation index used to distinguish between pulses carrying the bit 0 and the bit 1 for PPM scheme and d_j is the information symbol [3].

To eliminate catastrophic collisions due to multiple access, each user (indexed by k) is assigned a distinctive time shift pattern called a time hopping code. This provides an additional time shift of $c_i^{(k)} T_c$ seconds to jth monocycle in the pulse train. The modulation index δ can be chosen to optimize system performance. For performance prediction purposes, the data sequence $\left\{d_j^{(k)}\right\}_{j=-\infty}^{\infty}$ is modeled as a wide-sense stationary random process composed of equally likely symbols. In TH multiple access scheme, the "pulse repetition time" or "frame time" between two consecutive pulses is divided into a number of smaller time slots, the length of each slot being called as "chip duration". Code cardinality, denoted as N_h, is the total number of such time slots within each frame time. Moreover, each information bit is transmitted using N_S consecutive pulses, leading to (N_S, 1) repetition code. The resulting information bit rate is thus $R_b = \left(N_S\, T_f\right)^{-1}$. In this paper, the effect of code cardinality on BER performance of the system has been investigated and corresponding results have been presented. The effect of pulse repetition rate, N_S for different code cardinality values is also shown. The simulations have been performed for a typical 16.6 Mbit/s indoor UWB system, however, the nature of how the system performance would deviate in case of increased data rate requirements is also indicated by showing results for 50 Mbit/s and 100 Mbit/s TH-PPM systems.

The paper is organized as follows: Section 2 provides a brief description of the specific pulse shape used. The discussion is followed by Section 3 describing briefly the system model. Results obtained in this investigation have been presented in Section 4. Finally, Section 5 concludes the paper.

2. Ultra Wideband Pulse Shape

The selection of the most appropriate pulse shape for UWB transmission requires that it match the Federal Communications Commission (FCC) regulated power spectral density (PSD) mask in a reasonably good manner and at the same time increases signal bandwidth. Considering the nth derivative of Gaussian pulse as the transmitted pulse of UWB transmission, where $A_{\max}$ is the peak power spectral density that has been set as limit by the FCC in USA, the PSD of the transmitted signal can be expressed as [4]

$$|P_t(f)| \equiv A_{\max}|P_n(f)| = \frac{A_{\max}(2\pi f\sigma)^{2n}\mathrm{e}^{\{-(2\pi f\sigma)^2\}}}{n^n\mathrm{e}^{(-n)}}. \quad (2)$$

Based on normalized PSD of *n*th order Gaussian derivative pulse, applying bisection method as shown in [4], the fourth order derivative of Gaussian pulse with pulse shaping factor of 0.168 ns has been chosen in our simulations. This is because as shown in Fig. 1 this specific condition fulfills the FCC regulated PSD mask in the most appropriate manner. Figure 1 also presents the effect of higher order Gaussian derivative pulses on FCC regulated PSD mask for UWB indoor communications. The basic fourth order derivative of Gaussian pulse has been illustrated in Fig. 2.

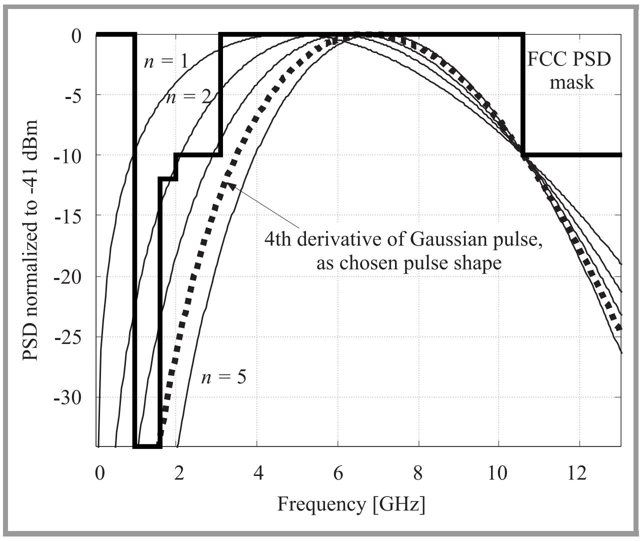


Fig. 1. PSD of 4th order derivative of Gaussian pulse fulfilling indoor UWB PSD requirement for indoor systems.

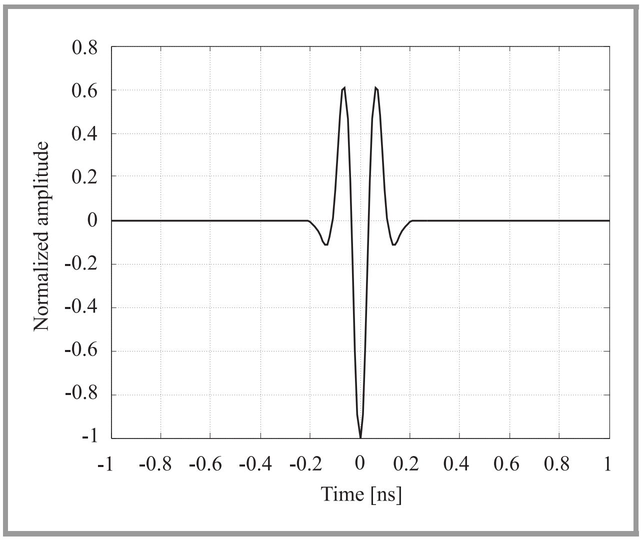


Fig. 2. Fourth order derivative of Gaussian pulse.

3. System Model

In this paper, the effects of different code cardinality values on bit error rate have been studied with the most recent IEEE 802.15.3a [5] UWB channel model. The overall system model is shown in Fig. 3 and is described in detail in the following.

3.1. Transmitter Section

Time hopping multiple access has been used along with PPM modulation scheme in the transmitter. The transmitter specifications along with the remaining parameters used in simulation of the complete system have been presented in Table 1. The fourth derivative of Gaussian pulse with a power decay factor of 0.168 ns best satisfies the FCC regulated PSD mask and so it has been used in this paper.

Table 1
Parameters used in simulation of the complete system

Parameter	Value used in simulation
Source data rate	16.6 Mbit/s
Processing gain	20.79 dB
Average transmitter power	–30 dBm
Sampling frequency	30 GHz
No. of pulse per bit	1, 2, 4, 8
Frame time	60.1 ns
Periodicity of the TH code	2000
Chip time	1 ns
Multi-user interference	Single user and multiple user scenarios
Receiver	Selective RAKE, 8 arms
Channel model	IEEE 802.15.3a, CM-3: 4–10 m, NLOS
Modulation scheme	PPM
Multiple access	TH
Time shift for PPM	0.5 ns
Pulse shape	Gaussian 4th derivative
Pulse width	0.5 ns
Pulse decay factor	0.168 ns
Code cardinality	5, 10, 15, 20, 30, 50

3.2. IEEE 802.15.3a UWB Channel Model

In IEEE 802.15.3a UWB multipath channel model, a modified Saleh-Valenzuela model has been adopted on the basis of observed clustering phenomenon in several channel measurements. Log-normal distribution rather than a Rayleigh distribution for the multipath gain magnitude has been recommended. In addition, independent fading is assumed for each cluster as well as each ray within the cluster. There-

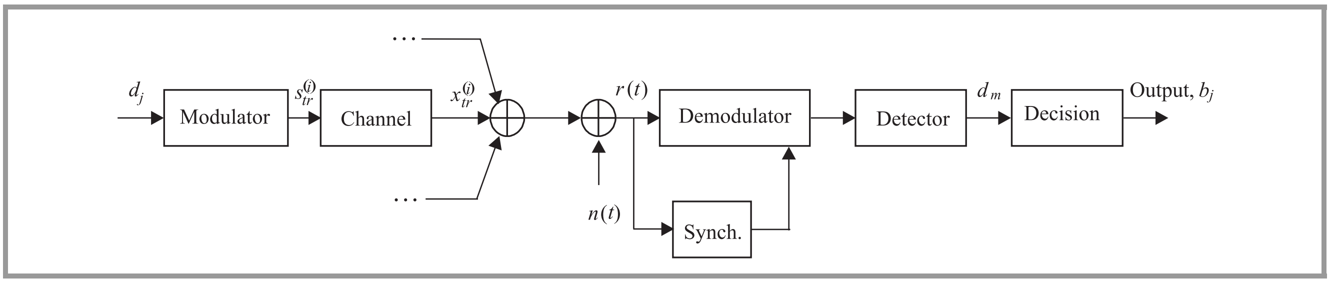


Fig. 3. The system model.

fore, the discrete-time impulse response of the multipath channel model can be described as [6]

$$h_i(t) = X_i \sum_{l=0}^{L} \sum_{k=0}^{K} \alpha_{k,l}^{i} \delta\left(t - T_l^i - \tau_{k,l}^{i}\right), \tag{3}$$

where $\{\alpha_{k,l}^i\}$ are the multipath gain coefficients, $\{T_l^i\}$ is the delay of the lth cluster, $\{\tau_{k,l}^i\}$ is the delay of the kth multipath component relative to the lth cluster arrival time (T_l^i), $\{X_i\}$ represents the log-normal shadowing and i refers to the ith realization.

So, according to the model, T_l represents the arrival time of the first path (ray) of the lth cluster; $\tau_{k,l}$ is the delay of the kth path (ray) within the lth cluster relative to the first path arrival time, T_l. The IEEE 802.15.3a channel model has been already explained in several available literature, e.g., in [6], for four special cases depending on transmitter to receiver distance and the availability of line of sight (LOS) path between them. In this paper, CM-3 (4–10 m, NLOS) condition of the channel has been used because CM-3 represents a typical indoor environment in the transmitter-receiver distance of 4–10 m. The corresponding channel parameters are shown in Table 2.

Table 2
Model parameters in IEEE 802.15.3a UWB channel [6]

Model parameters	CM-3 NLOS at 4–10 m
Cluster arrival rate, Λ [1/ns]	0.0667
Ray arrival rate, λ [1/ns]	2.1
Cluster decay factor, Γ	14
Ray decay factor, γ	7.9
Std. dev. of cluster log-normal fading, σ_1 [dB]	3.3941
Std. dev. of ray log-normal fading, σ_2 [dB]	3.3941
Std. dev. of total multipath log-normal fading, σ_x [dB]	3

3.3. Receiver Section

In the receiver section, selective RAKE receiver has been used. The received signal is the sum of replicas of the transmitted signals. The received signal is, therefore, expressed as

$$r(t) = X \sum_{l=1}^{L} \sum_{k=1}^{K} \alpha_{k,l}\, s_{tr}(t - T_l - \tau_{k,l}) + (n(t) + n_f(t), \tag{4}$$

where $s_{tr}(t)$ is the transmitted signal which suffers from attenuation and time delay in multipath propagation, $n(t)$ is zero mean AWGN (additive white Gaussian noise) and $n_f(t)$ is the multiple user interference signal. The remaining symbols are as described in Subsection 3.2.

For simulation of this study RAKE receiver with 8 arms has been chosen. This is because in our simulation 8 RAKE arms have shown to give better results providing a trade-off between number of RAKE arms and desired BER of the system. First arm is locked to the first multipath component, m_1. Multipath component, m_2 arrives τ_1 time units later than m_1 and is captured and so on. All decision statistics are weighted by a weighting factor, α to form overall decision statistics. The signals are then integrated over the entire period. The integrated signal is then compared with the appropriate threshold value to receive the better estimate of the transmitted signal. Hard decision detection (HDD) has been chosen at the receiver section because in our simulation results HDD has been found to be more efficient than soft decision detection (SDD) for TH-PPM UWB systems.

4. Results

Code cardinality (N_h) has a significant effect on the system bit error rate in case of TH-PPM UWB system using IEEE 802.15.3a UWB channel model. Code cardinality influences BER performance at the receiver quite significantly if its value is as large as in the range from 30 to 50. As shown in Fig. 4 the system BER performance is almost unchanged when code cardinality value is low, e.g., in the range from 5 to 20. In this range BER performance is almost independent of the code cardinality for up to an E_b/N_o value of 12 dB, but beyond that point the BER performance degrades for higher code cardinality values. For instance, the condition of $N_h = 20$ gives comparatively worse performance compared to that of N_h in the range from 5 to 15. On the other hand, our simulation results show that when code cardinality value (N_h) is set at 30, the BER performance degrades significantly. Also, if code cardinality is

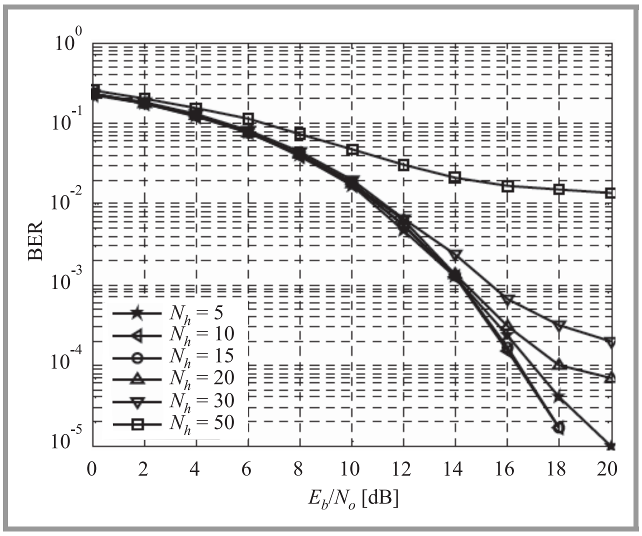


Fig. 4. BER performance for various code cardinality values, at a distance between transmitter and receiver of 5 m, with 8 selective RAKE arms, one pulse per information bit, $N_S = 1$, no multiple user interference.

again increased to 50 the performance is the worst of all. It is important to note that all of the above cases have been simulated for a TH-PPM system of a data rate capacity of 16.6 Mbit/s with the number of pulses to represent one bit (N_S) set as $N_S = 1$. For our study, the frame time (T_S) of the TH-PPM transmitter has been chosen as 60.1 ns, which makes the overall data rate 16.6 Mbit/s. The chosen values of N_S and T_S pair determine and limit the maximum number of elements in the code cardinality vector, i.e., the value of code cardinality (N_h) in TH-PPM transmitter section. In view of the above discussions it is evident that the effect of code repetition rate (N_S) and the frame time (T_S) should also be investigated along with the effects of code cardinality on BER performance of TH-PPM system.

Figure 5 shows the effect of code repetition rate (N_S) of TH-PPM transmitter on overall system BER performance. Again, the system is considered as a 16.6 Mbit/s TH-PPM system with a constant value of code cardinality set at $N_h = 5$. In order to keep the data rate constant all through, the chosen values of (N_S, T_S) pair as used to generate the results in Fig. 5 have been illustrated in Table 3.

Table 3
Chosen values of (N_S, T_S) pair for Fig. 5

Code repetiition rate, N_S	Frame time, T_S [ns]	Data rate = $1/(N_S T_S)$ [Mbit/s]
1	60.1	16.6
2	30	16.6
4	15	16.6
8	7.5	16.6

As presented in Fig. 5 at a code cardinality value of $N_h = 5$ for a 16.6 Mbit/s TH-PPM UWB system, higher code repetition rates result in severe performance degradation. For example, for a desired BER of $2 \cdot 10^{-2}$, one pulse per information bit condition (i.e., $N_S = 1$) provides E_b/N_o gains of 5 dB, 6.5 dB and 7.5 dB over the performances of $N_S = 2$, $N_S = 4$ and $N_S = 8$ cases, respectively. However, this is also important to note that these results are not the results for the effects of data rate on BER of TH-PPM system, which is usually done by varying any or both of code repetition rate, N_S and frame time, T_S.

As mentioned earlier, the results presented so far have been obtained from simulations based on a 16.6 Mbit/s TH-PPM UWB system. However, the effect of higher system data rate requirements on BER performance has also been shown in Fig. 6. It is clearly illustrated in Fig. 6 that if the system data rate requirement is increased from 16.6 Mbit/s to 50 Mbit/s or 100 Mbit/s, the BER performance degrades

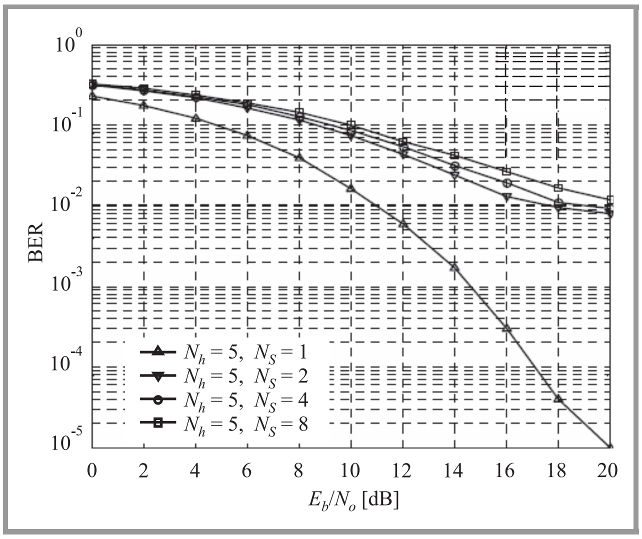


Fig. 5. BER performance for different code repetition rates at distance between transmitter and receiver of 5 m, 8 selective RAKE receiver arms, code cardinality, $N_h = 5$, no multiple user interference.

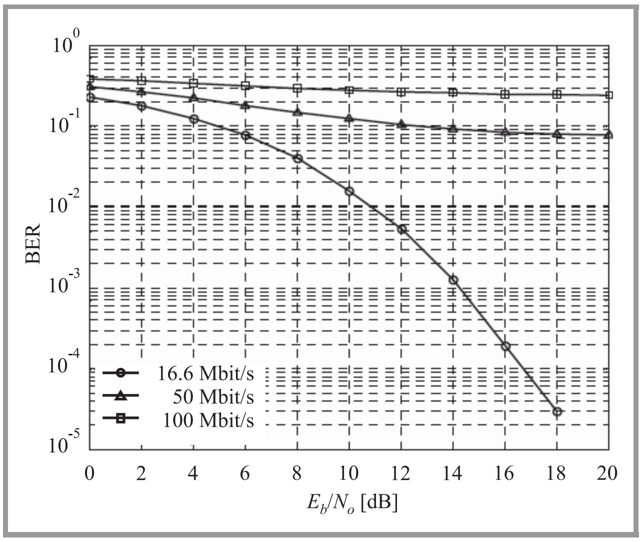


Fig. 6. BER performance of TH-PPM UWB system at different system data rate requirements, at distance between transmitter and receiver of 5 m, 8 selective RAKE receiver arms, no multiple user interference.

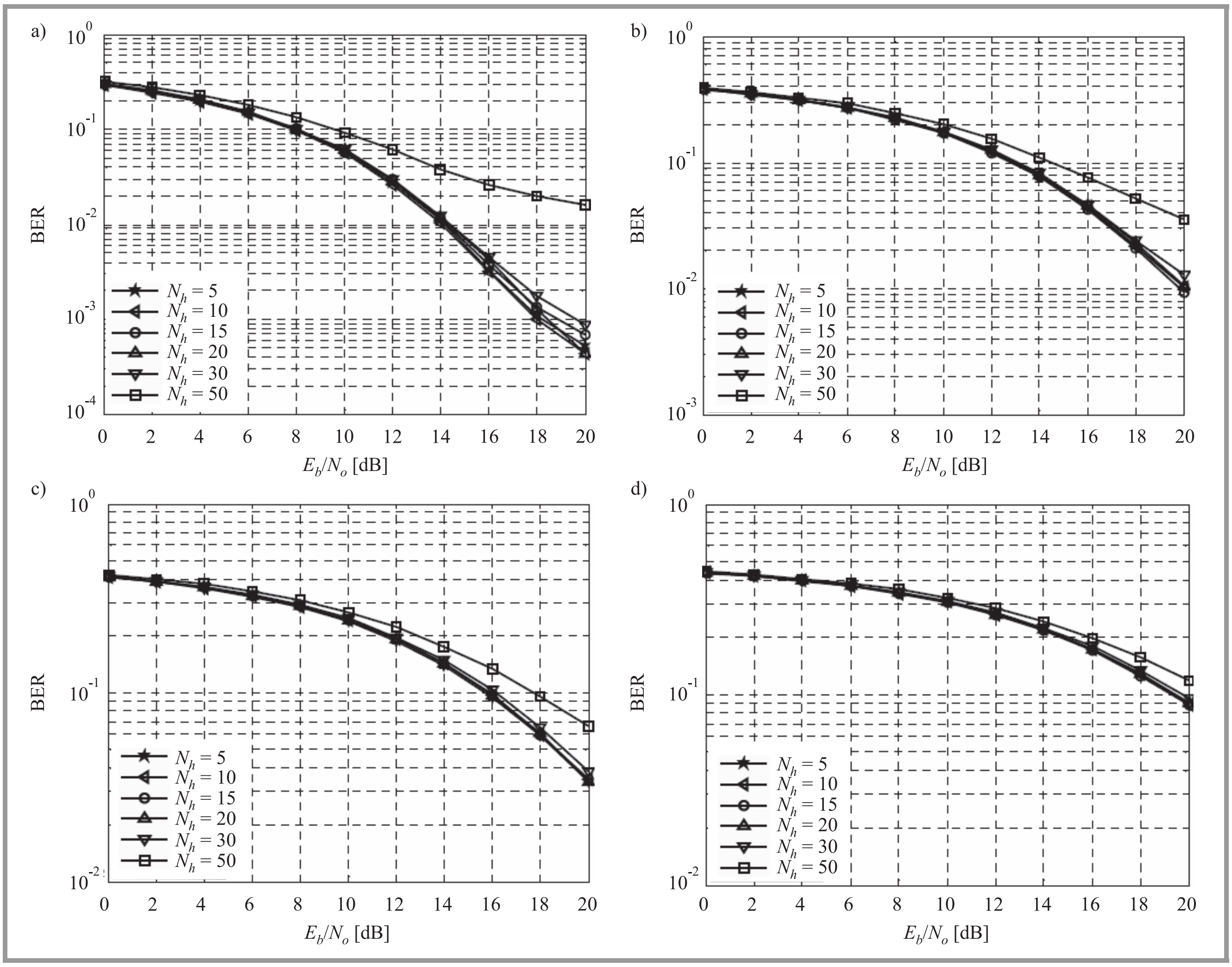


Fig. 7. BER performance of TH-PPM UWB systems with: (a) 1; (b) 5; (c) 10; (d) 20 interfering users for various code cardinality values, at a distance between transmitter and receiver of 5 m, with 8 selective RAKE arms, one pulse per information bit, $N_S = 1$.

enormously. In view of the simulation results so far obtained in our work, a code cardinality value of 10 to 15 is recommended for a 16.6 Mbit/s TH-PPM UWB system with no multiple user interference, because such a condition provides the best BER performance for TH-PPM UWB indoor communication system.

Moreover, system BER performance for different multiple user scenarios of 1, 5, 10 and 20 interfering users has also been shown in Fig. 7. Results in Fig. 7 suggest that BER performance is almost the same for all code cardinality values unless it is as high as $N_h = 50$. As a result, a code cardinality value of 10 to 15 can also suit well for TH-PPM UWB systems with multiple user interference. It is also noted that TH-PPM UWB system suffers serious performance degradations in the presence of multiple user interference.

5. Conclusions

In this paper, the effect of code cardinality on the BER performance of TH-PPM UWB communication system has been presented. TH-PPM scheme has been chosen to investigate into this effect because TH-PPM is the most widely used scheme for impulse radio systems. The simulations in this study considered both single and multiple user scenarios. Separate results have also shown the effect of pulse repetition rates along with those of code cardinality. This is because in simulating the effect of code cardinality only one pulse per bit transmitted has been used and so investigation into multiple pulses per bit transmitted is worth-considering in TH-PPM UWB systems. Another important implication of investigating into code cardinality in case of TH-PPM UWB systems is that the PSD of the TH-PPM transmitter signal is related to code cardinality. That is why an appropriate analysis of code cardinality and its influence on BER of the system needs much research attention.

This paper concludes that a code cardinality value in the range from 10 to15 can be recommended as the most appropriate code cardinality value for 16.6 Mbit/s TH-PPM UWB systems with both single and multiple user scenarios, especially in view of BER performance. However, for other systems the code cardinality value must be chosen giving much attention to code repetition rate and frame time as well as to the data rate requirement of the system.

References

[1] N. Laurenti, T. Erseghe, and V. Cellini, "On the performance of TH-PPM and TH-PAM as transmission formats for UWB communications", in *Proc. IEEE Veh. Technol. Conf. VTC – Spring*, Milan, Italy, 2004.

[2] J. Zhang, R. A. Kennedy, and T. D. Abhayapala, "Conditions and performance of ideal RAKE reception for UWB signals in lognormal fading channels", *Int. J. Wirel. Inform. Netw.*, vol. 10, no. 4, pp. 193–200, 2003.

[3] M. Z. Win and R. A. Scholtz, "Impulse radio: how it works", *IEEE Commun. Lett.*, vol. 2, no. 2, pp. 36–38, 1998.

[4] H. Sheng, P. Orlik, A. M. Haimovich, L. J. Cimini Jr., and J. Zhang, "On the spectral and power requirements for ultra-wideband transmission", in *Proc. Int. Conf. Commun. ICC*, Anchorage, Alaska, USA, 2003, pp. 738–742.

[5] "IEEE Standard for information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements". Part 15.3: "Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPANs)", IEEE Std 802.15.3$^{\text{TM}}$-2003.

[6] J. R. Foerster, M. Pendergrass, and A. F. Molisch, "A channel model for ultrawideband indoor communication", in *Proc. Wirel. Pers. Multimed. Commun. WPMC-03*, Kanagawa, Japan, 2003, vol. 2, pp. 116–120.

Mohammad Upal Mahfuz received his B.Sc. engineering degree in electrical and electronic engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 2002 and Master of engineering degree in telecommunications from Asian Institute of Technology (AIT), Thailand, in 2005. Currently, he is working towards his graduate studies in the Department of Geomatics Engineering at University of Calgary, Canada. His current research interests include ultra wideband wireless communications, mobile communications and satellite-based positioning and navigation systems.

e-mail: upal41@yahoo.com
Department of Geomatics Engineering
University of Calgary
2500 University Drive NW
Calgary, Alberta, T2N 1N4 Canada

Kazi M. Ahmed received the M.Sc. Eng. degree in electrical engineering from the Institute of Communications, Leningrad, USSR, and the Ph.D. degree from the University of Newcastle, NSW, Australia, in 1978 and 1983, respectively. Currently, he is a Professor of telecommunications in the School of Engineering and Technology, Asian Institute of Technology, Pathumthani, Thailand. His current research interests include digital signal processing, antenna array processing, tropospheric and ionospheric propagation studies for microwave, very high frequency-ultra high frequency communications, and satellite communications.

e-mail: kahmed@ait.ac.th
Telecommunications Program
Asian Institute of Technology (AIT)
PO Box: 4, Khlongluang
Pathumthani 12120, Thailand

Nandana Rajatheva received the B.Sc. degree in electronic and telecommunication engineering (with first class honors) from the University of Moratuwa, Sri Lanka, and the M.Sc. and Ph.D. degrees from the University of Manitoba, Winnipeg, Canada, in 1987, 1991, and 1995, respectively. Currently, he is an Associate Professor of telecommunications in the School of Engineering and Technology, Asian Institute of Technology, Pathumthani, Thailand. Earlier, he was with the University of Moratuwa, Sri Lanka, where he became a Professor of electronic and telecommunication engineering in June 2003. From May 1996 to December 2001, he was with TC-SAT as an Associate Professor. His research interests include mobile and wireless communications, coding and modulation techniques, space time processing for multiple input-multiple output systems, and communication theory.

e-mail: rajath@ait.ac.th
Telecommunications Program
Asian Institute of Technology (AIT)
PO Box: 4, Khlongluang
Pathumthani 12120, Thailand

Paper

High-Frequency Power Amplitude Modulators with Class-E Tuned Amplifiers

Juliusz Modzelewski and Mirosław Mikołajewski

Abstract—A high-frequency power amplifier used in a drain amplitude modulator must have linear dependence of output HF voltage V_o versus its supply voltage V_{DD}. This condition essential for obtaining low-level envelope distortions is met by a theoretical class-E amplifier with a linear shunt capacitance of the switch. In this paper the influence of non-linear output capacitance of the transistor in the class-E amplifier on its $V_o(V_{DD})$ characteristic is analyzed using PSPICE simulations of the amplifiers operating at frequencies 0.5 MHz, 5 MHz and 7 MHz. These simulations have proven that distortions of the $V_o(V_{DD})$ characteristic caused by non-linear output capacitance of the transistor are only slight for all analyzed amplifiers, even for the 7 MHz amplifier without the external (linear) shunt capacitance. In contrast, the decrease of power efficiency of the class-E amplifier resulting from this effect can be significant even by 40%.

Keywords*— *high-efficiency amplitude modulators, non-ZVS operation, optimum operation, PSPICE simulations, sub-optimum operation.

1. Introduction

Power amplitude modulators are commonly used as the output stage of broadcasting and radio-communication transmitters with amplitude modulation (AM) as well as in special high-efficiency linear amplifiers utilizing the envelope elimination and restoration (EER) method.

In the basic power AM modulator the amplitude of its high-frequency (HF) output signal is modulated by varying the supply voltage of a HF power amplifier. This method of amplitude modulation is called the drain (collector or anode) modulation. The HF power amplifier applied in the drain modulator must fulfil two important requirements. First of all, to avoid non-linear distortions of the output-signal envelope, the amplitude V_o of the output voltage in this amplifier must be directly proportional to its supply voltage V_{DD}:

$$V_o = k \cdot V_{DD}\,, \tag{1}$$

where $k = \text{const.}$, $v_o(t) = V_o \sin(2\pi\, f_c\, t)$ is the output voltage, f_c is the carrier frequency. The relationship $V_o(V_{DD})$ of the amplifier is called the amplitude-modulation characteristic. Another highly desirable feature of such the AM-modulated amplifier is its high efficiency in the whole range of output voltage $0 - V_{o\max}$ to decrease power losses in the AM modulator. Thus, cost, dimensions, and weight of the modulators can be reduced, which is very important particularly in mobile, battery-operated transmitters. Therefore, the dependence of efficiency on the supply voltage $\eta(V_{DD})$ is also a very important characteristic of the AM-modulated HF amplifier.

Similar problems arise in industrial electronics when HF power regulation is needed, e.g., in inductive or dielectric heaters. In these applications high-efficiency is a basic feature of the regulated power amplifier. The linearity of $V_o(V_{DD})$ is desirable but not required.

The class-E tuned power amplifier (Fig. 1a) with its high efficiency as well as a highly linear amplitude modulation characteristic theoretically satisfies the condition (1), which makes it an attractive choice for a high-efficiency amplitude-modulated transmitter. However, in the real-world class-E amplifiers there are phenomena disturbing the drain amplitude modulation [1], [2]. Above all, the quality of the amplitude modulation by varying the supply voltage deteriorates with the increase of the modulating signal frequency. It results from the fact that, firstly, the large (theoretically infinite) inductance of the RF power-supply choke L_{CH} limits the available rate of change of the output-HF-voltage envelope. Therefore the modulation depth m decreases with increase of the modulating-signal frequency (linear distortion). Secondly, the envelope of the output HF signal is non-linearly distorted because the lower and upper sideband components of the AM signal are transmitted by the series resonant branch L_{sr}, C_{sr}, R_L (Fig. 1) with different gains and phase shifts. This effect is caused by the fact that the branch L_{sr}, C_{sr} is not in resonance at the carrier frequency f_c and its impedance is inductive. The non-linear distortions become worse with the increase of the loaded quality factor of the L_{sr}, C_{sr}, R_L branch, with the increase of the ratio of the modulating-signal bandwidth to the carrier frequency f_c, and with the depth of modulation m [3]. These distortions can be high, up to 12% [1].

The non-linear distortions of the envelope can be also caused by non-linearities of the static characteristic of the drain modulation $V_o(V_{DD})$ in real class-E amplifiers. For small values of the supply voltage V_{DD} the non-linearity of $V_o(V_{DD})$ is caused by direct transmission of the input HF signal to the output by the reverse drain-gate capacitance C_{rss} of the transistor [3]. Hence, the amplitude V_o is non-zero for $V_{DD} = 0$. This effect increases with the operating frequency of the amplifier.

For a high operating frequency the non-linearity of the amplifier characteristic $V_o(V_{DD})$ can be also caused by the non-linear output capacitance C_{oss} of the transistor because this capacitance forms most of the shunt capacitance

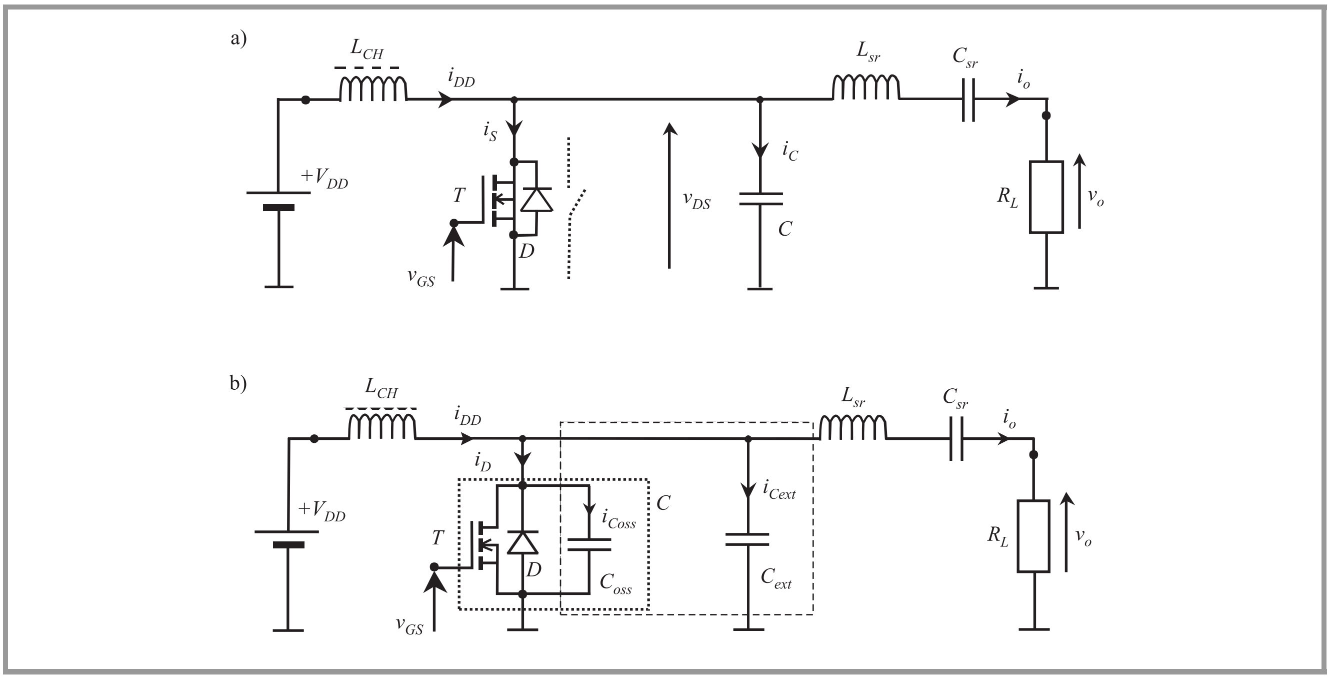


Fig. 1. Class-E tuned power amplifier: (a) basic circuit; (b) components of the shunt capacitance of the switch ($C = C_{ext} + C_{oss}$).

$C = C_{ext} + C_{oss}$ in the class-E amplifier (Fig. 1b). Then the variations of the supply voltage V_{DD} cause substantial changes in the value of C, which increases considerably with decreasing drain-source voltage (Fig. 2). This modifies in a significant way the operation of the class-E amplifier because both decrease and increase of the supply voltage mistune this amplifier changing shapes of the drain voltage and current waveforms. Hence, the function $V_o(V_{DD})$ of the class-E amplifier without external shunt capacitance may be non-linear. Besides, for certain ranges of V_{DD} the class-E amplifier may not operate optimally and its efficiency become worse. Moreover, the resonant frequency of amplifier parallel resonant circuit $C - L_{sr} - C_{sr} - R_l$ also changes with V_{DD} causing unwanted amplitude and phase modulation of the amplifier output signal.

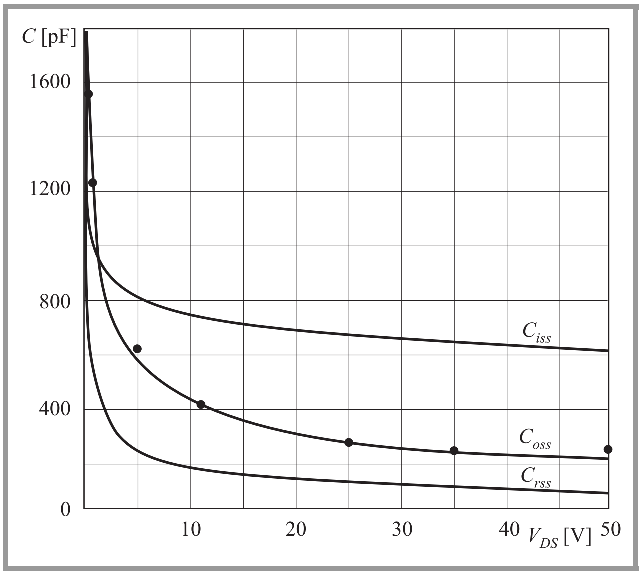


Fig. 2. Capacitances C_{iss}, C_{oss}, C_{rss} of transistor IRF530 versus drain-to-source voltage [4] and the values of C_{oss} extracted from the modified PSPICE model (●).

The aim of this paper is to analyze effects of non-linearity of the transistor output capacitance C_{oss} for the drain-modulation static characteristic $V_o(V_{DD})$ and drain-efficiency static characteristic $\eta_D(V_{DD})$ of the HF class-E amplifier. It is necessary to compare these characteristics determined for amplifiers with the same transistor but operating at different frequencies, i.e.:

– low-frequency amplifier with approximately linear shunt capacitance ($C_{ext} \gg C_{oss}$);

– medium-frequency amplifier with noticeably non-linear shunt capacitance ($C_{ext} \approx C_{oss}$);

– high-frequency amplifier with maximally non-linear shunt capacitance ($C_{ext} \approx 0$).

This analysis can be done by PSPICE (Personal Computer Simulation Program with Integrated Circuit Emphasis) simulations. In the analyzed class-E amplifiers the transistor IRF530 has been applied.

2. Modes of Operation of the Class-E Amplifier

The class-E amplifier is a high-efficiency switch-mode resonant amplifier. Its high efficiency results from very much

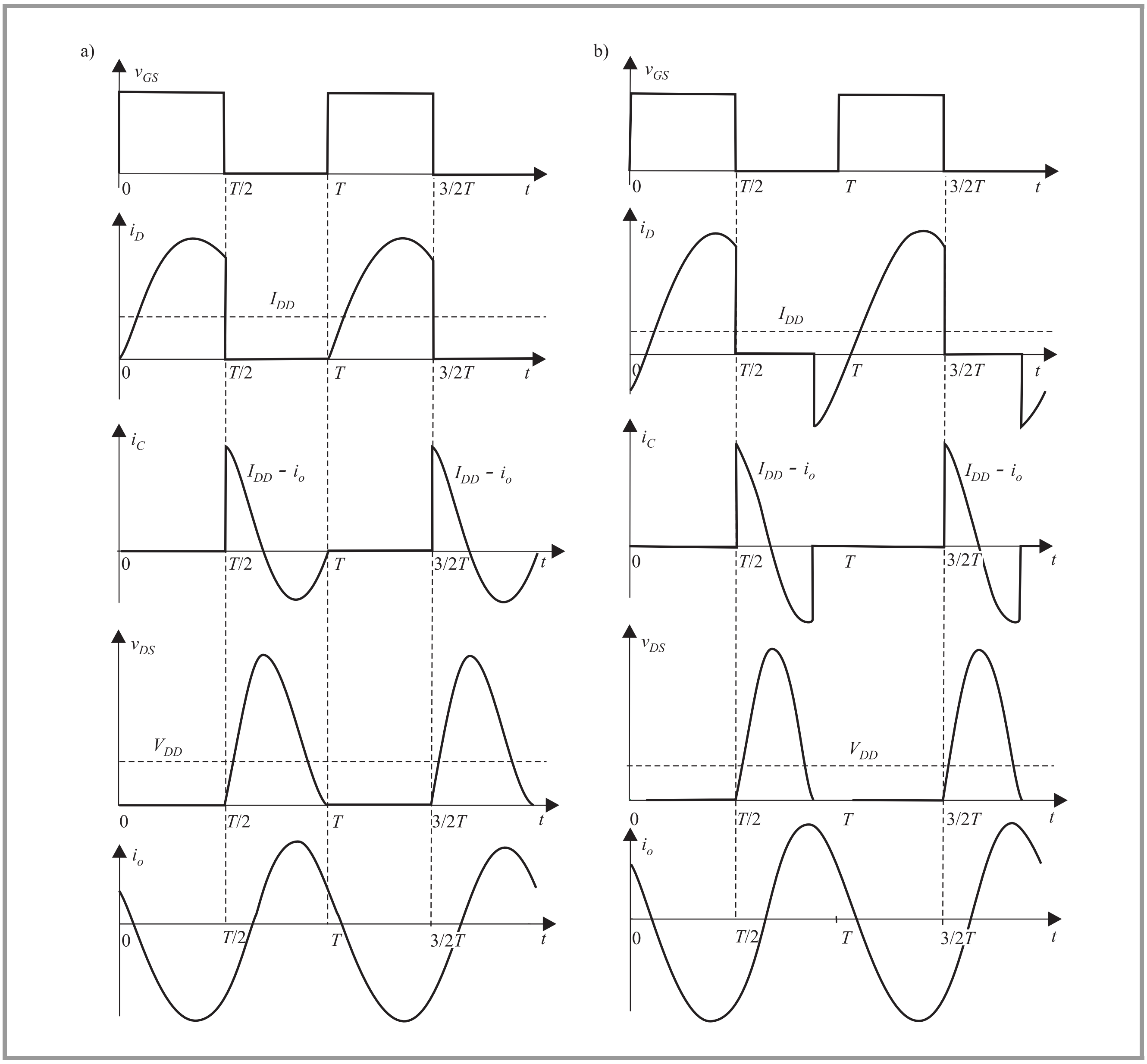


Fig. 3. Optimum (a) and sub-optimum (b) operation of the class-E amplifier.

reduced power losses in the transistor. To achieve high efficiency, firstly, the transistor in the amplifier operates as a switch to reduce power losses caused by its conducted current. Secondly, switching losses resulting from a finite transition time between on and off states of the transistor switch are also reduced. The decrease of switching losses is obtained by shaping the drain current and drain voltage waveforms of the transistor by the resonant circuit in the amplifier to achieve so-called zero-voltage switching (ZVS) and/or zero-current switching (ZCS) operation of the transistor switch. The ZVS and/or ZCS operation means that at the switching moments the drain-source voltage of the transistor is zero and/or its drain current is zero as well. Then the instantaneous value of drain power loss in the transistor switch at its switching moments is zero or is very much reduced.

The class-E amplifier (Fig. 1) consists of a choke L_{CH} supplying a DC current I_{DD} to the circuit, transistor switch T and a parallel-series resonant circuit $C-L_{sr}-C_{sr}$ with load resistance R_L. When the switch T is on it conducts the supply current I_{DD} and a sinusoidal current i_o of the series resonant circuit $L_{sr}-C_{sr}-R_L$. Then the power losses in the transistor depend on its conducted current $i_D = I_{DD} - i_o$ and the transistor on-resistance. When the switch is off, the current $I_{DD} - i_o$, at first charges (for $I_{DD} - i_o > 0$) and then discharges (for $I_{DD} - i_o < 0$) the parallel capacitance C shaping the v_{DS} waveform. The maximum value of v_{DS} occurs for $I_{DD} - i_o = 0$.

In the operation of the amplifier three different modes can be identified, which are presented in Figs. 3 and 4 [5]. Figure 3a shows the waveforms in the class-E amplifier for its optimum operation. The switch T turns on at zero drain voltage and zero drain current (i.e., $v_{DS}(t=0)=0$ and $Cdv_{DS}/dt\big|_{t=0} = i_D(t=0)=0$). Therefore there is no switching loss at the transistor turn on. When the switch T turns off there is a jump change in the drain current i_D, but the drain voltage rises slowly from zero, which results in low switching losses. This mode is characterized by the highest efficiency (practical values of 98% are achieved), and the amplifier is usually designed to work in this mode. The maximum value of drain current and drain voltage are then typically $I_{D\max} = 2.78\, I_{DD}$, $V_{DS\max} = 3.61\, V_{DD}$, respectively (for loaded quality factor of the $L_{sr}-C_{sr}-R_L$ circuit equal to $Q_l = 5$ [6]).

Waveforms for the sub-optimum operation of the class-E amplifier are given in Fig. 3b. In this mode of operation the shunt capacitor C is discharged to zero before transistor T is turned on by the driving signal v_{GS}. Then the voltage v_{DS} becomes negative and the anti-parallel diode D of the transistor turns on conducting the negative current $i_D = I_{DD} - i_o$ and maintaining v_{DS} voltage close to zero till the turn-on instance of the transistor. Thus, the transistor T turns on and off at ZVS and non-ZCS conditions. The use of diode D allows avoiding a significant negative drain voltage appearing across the transistor, which would cause high turn-on switching losses. The amplifier efficiency in the sub-optimum mode is lower then in the optimum mode but still can be high up to 95%. This operation mode can be obtained by decreasing load resistance R_l. The maximum values of drain current and drain voltage in the sub-optimum mode are higher then in the optimum mode ($V_{DS\max}$ can 4.6 times exceed V_{DD}) even though the output power is lower.

Figure 4 illustrates non-ZVS operation of the class-E amplifier (so-called non-optimum operation). In this mode of operation the transistor turns on when the drain voltage v_{DS} is non-zero. This results in high turn-on switching losses due to the current spike in the drain current i_D caused by rapid discharging of the shunt capacitor C. The efficiency in this mode of operation can be much decreased and power losses in the transistor can be high requiring proper cooling of the transistor if one expects the non-ZVS mode to occur in the amplifier operation. This mode of operation of the class-E amplifier is be obtained by, e.g., increasing load resistance above its optimum-mode value. The maximum values of $I_{D\max}$ and $V_{D\max}$ as well as the output power are lower than in the optimum mode.

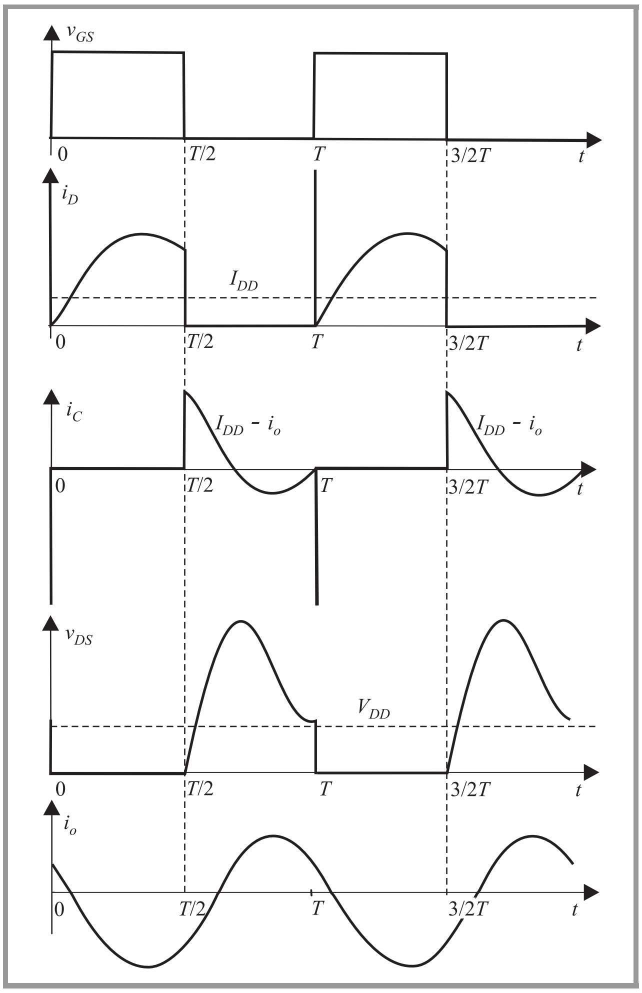


Fig. 4. Non-ZVS (non-optimum) operation of the class-E amplifier.

3. Model of the Output Capacitance of Power MOSFETs

The parasitic output capacitance C_{oss} of power MOSFETs is mainly the capacitance of the reverse biased p-n junction body diode. Therefore the small-signal value of C_{oss} can be expressed by:

$$C_{oss} \cong C_{jO}\left(1+\frac{v_{DS}}{V_{BI}}\right)^{-MJ}, \tag{2}$$

where: C_{jO} is the zero-bias capacitance, V_{BI} is the built-in potential of the body-diode junction, MJ is the grading coefficient of this junction ($MJ = 0.5$ for the step junction) [7]–[10]. In the basic PSPICE models of power MOSFETs [11] it is assumed $MJ = 0.5$ and value of C_{jO} is adjusted to obtain correct values of C_{oss} for medium values of v_{DS}. Thus, these models cannot be used for exact simulations of the class-E amplifier without external drain-source shunt capacitance of the transistor.

In [8] it was proven that the PSPICE model can describe correctly the power-MOSFET output capacitance as a function of v_{DS} if the grading coefficient is increased to $MJ = 0.77$. Therefore we modified the basic PSPICE model of the transistor IRF530 [11] assuming $MJ = 0.77$. For the drain-body junction zero-bias capacitance increased to $C_{bd} = 2$ nF (instead of 1.151 nF) the function $C_{oss}(V_{DS})$ extracted from PSPICE model is approximately equal to the data published by the manufacturer [4] – see Fig. 2.

4. Analyzed Class-E Amplifiers

It was assumed that the analyzed low-, medium- and high-frequency class-E amplifiers have the following parameters:

- switch duty cycle $D = 0.5$;
- maximum supply voltage $V_{DD\max} = 12$ V;
- maximum output power $P_{o\max} = 12$ W;
- loaded quality factor of the series resonant circuit $Q_1 = 5$.

For $Q_l = 5$ the load resistance of these amplifiers is equal to $R_L = 0.5249(V_{DD\max})^2/P_{o\max}$ [6]. Values of the resonant-circuit components (see Fig. 1a) for a given operating frequency f_o are: $L_{sr} = 5.673R_L/(2\pi f_o)$, $C_{sr} = 0.2269/(2\pi f_o R_L)$, $C = 0.2067/(2\pi f_o R_L)$ [6]. The operating frequency f_o of the low-, medium- and high-frequency class-E amplifiers with the chosen transistor IRF530 can be assumed as 0.5 MHz, 5 MHz, and 7 MHz, respectively. The calculated values of R_L and L_{sr}, C_{sr}, C are presented in Table 1.

Table 1
Components of the analyzed class-E amplifiers

f_o	L_{sr}	C_{sr}	C	R_L	L_{CH}	C_{ext}
[MHz]	[μH]	[nF]	[nF]	[Ω]	[μH]	[nF]
0.5	11.374	11.47	10.45	6.299	430	10.1
5	1.1374	1.147	1.045	6.299	43	0.497
7	0.8124	0.8193	0.7464	6.299	30.7	0.020

5. PSPICE Simulations

For PSPICE simulations it was assumed that the analyzed class-E amplifiers are driven by the 0.5 duty cycle square-wave generator consisting of the 10 V-peak-to-peak unipolar voltage source and the 3.5 Ω internal resistance. This source and resistance are an equivalent circuit of the integrated driver MIC 4423 applied in experimental amplifiers.

In the first step of simulation of the each amplifier (0.5 MHz, 5 MHz, and 7 MHz) the external shunt capacitance C_{ext} (Fig. 1b) was adjusted to ensure the optimum operation (i.e., with ZVS and ZCS turn on of the transistor IRF530) for $V_{DD} = 12$ V. The obtained values of C_{ext} are presented in Table 1. It should be noted that in the 7 MHz class-E amplifier (with the very low C_{ext} and nonlinear shunt capacitance of the switch) to ensure the optimum operation (Fig. 3a) it was necessary to adjust the duty cycle of the driving generator to $D = 52\%$ [12]. Inductance of the choke L_{CH} (Table 1) was chosen to obtain approximately constant supply current of the amplifiers. Drain voltage and current waveforms in the 0.5 MHz, 5 MHz, and 7 MHz amplifiers for $V_{DD} = 12$ V are presented in Figs. 5a, 7a, and 9a, respectively.

The carried out simulations have confirmed that for the sufficiently low supply voltage V_{DD} the each of the analyzed class-E amplifiers operates in the non-ZVS mode (Figs. 5b, 7b, and 9b). It results from the fact that the output capacitance of IRF530 strongly increases for low V_{DS} (Fig. 2). Obviously, in the 0.5 MHz amplifier this effect can be observed only at the very low supply voltage ($V_{DD} \leq 1$ V). In contrast, in the 7 MHz amplifiers

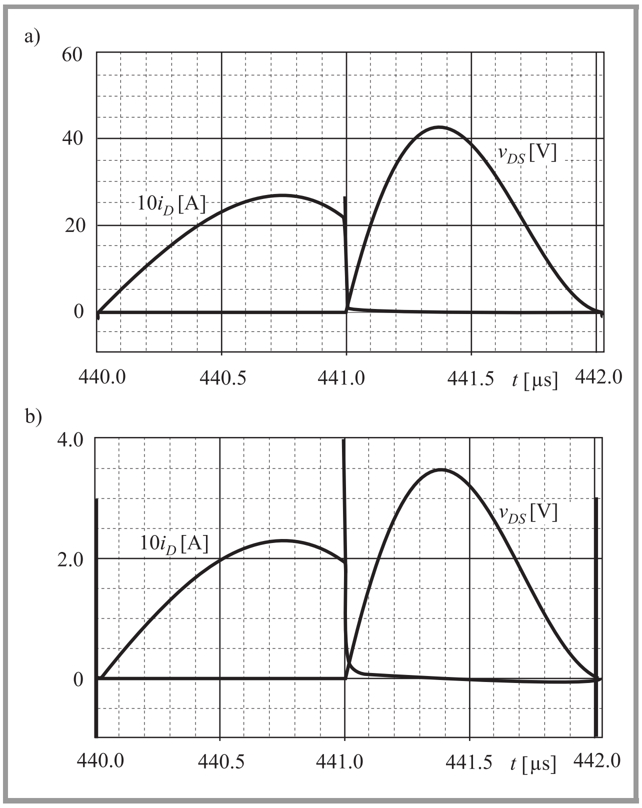


Fig. 5. Drain-source voltage v_{DS} and drain current i_D in the 0.5 MHz class-E amplifier (PSPICE simulation): (a) full supply voltage $V_{DD} = 12$ V; (b) reduced supply voltage $V_{DD} = 1$ V.

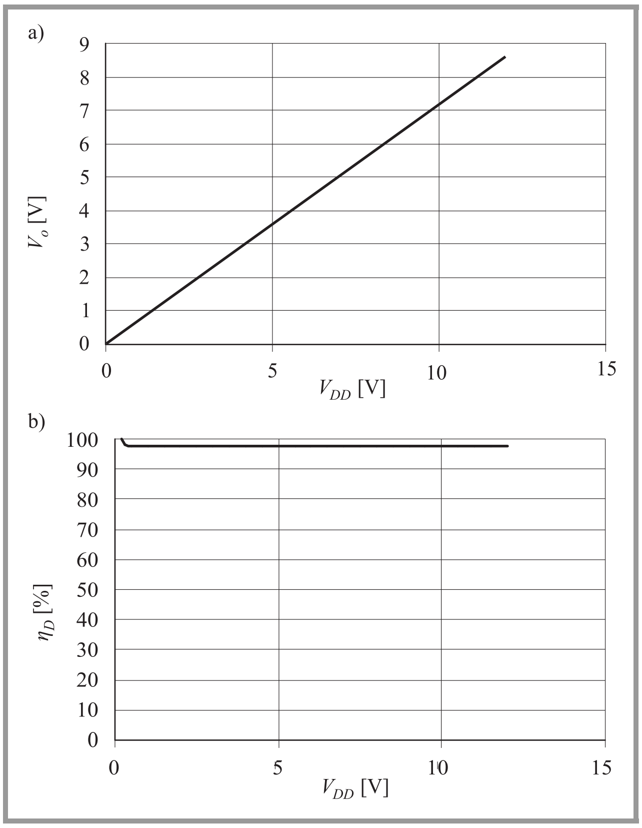


Fig. 6. Characteristic curves (a) $V_o(V_{DD})$ and (b) $\eta_D(V_{DD})$ of the 0.5 MHz class-E amplifier obtained by PSPICE simulations.

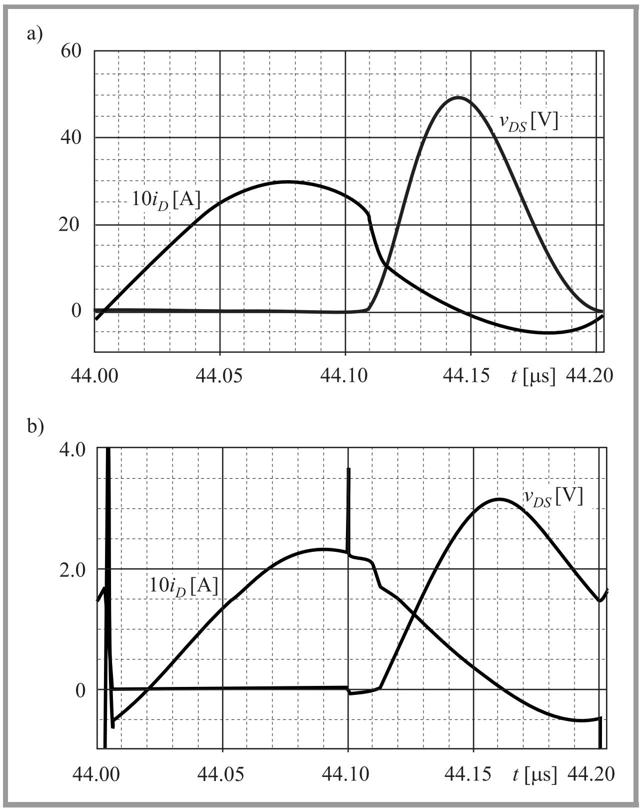


Fig. 7. Drain-source voltage v_{DS} and drain current i_D in the 5 MHz class-E amplifier (PSPICE simulation): (a) full supply voltage $V_{DD} = 12$ V; (b) reduced supply voltage $V_{DD} = 1$ V.

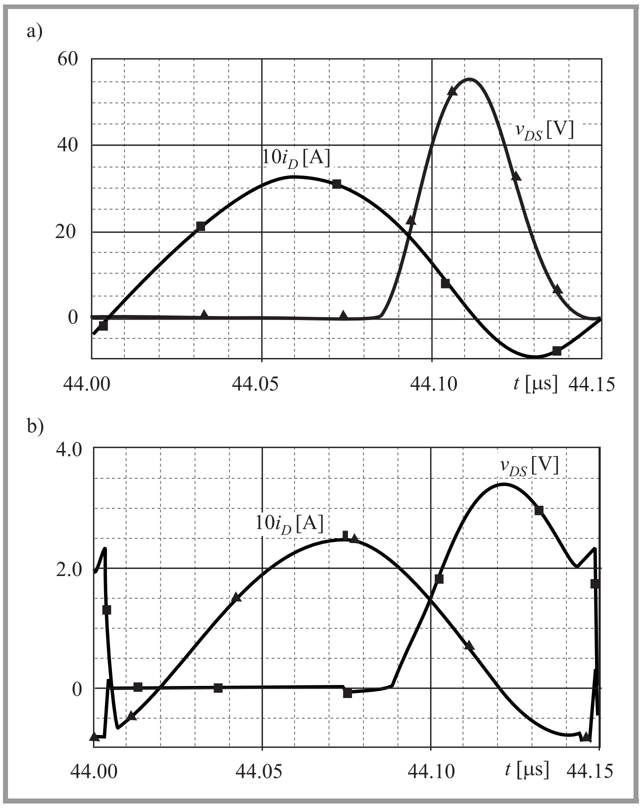


Fig. 9. Drain-source voltage v_{DS} and drain current i_D in the 7 MHz class-E amplifier (PSPICE simulation): (a) full supply voltage $V_{DD} = 12$ V; (b) reduced supply voltage $V_{DD} = 1$ V.

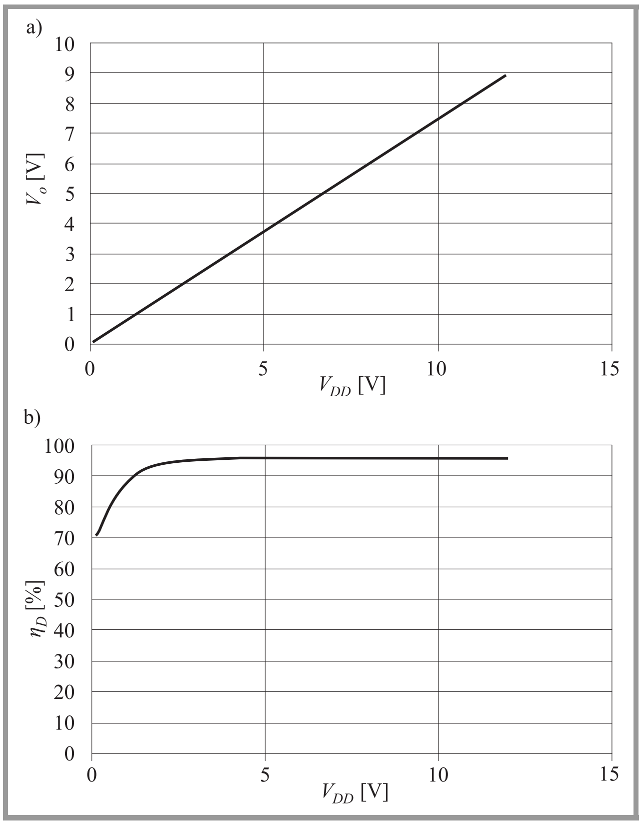


Fig. 8. Characteristic curves (a) $V_o(V_{DD})$ and (b) $\eta_D(V_{DD})$ of the 5 MHz class-E amplifier obtained by PSPICE simulations.

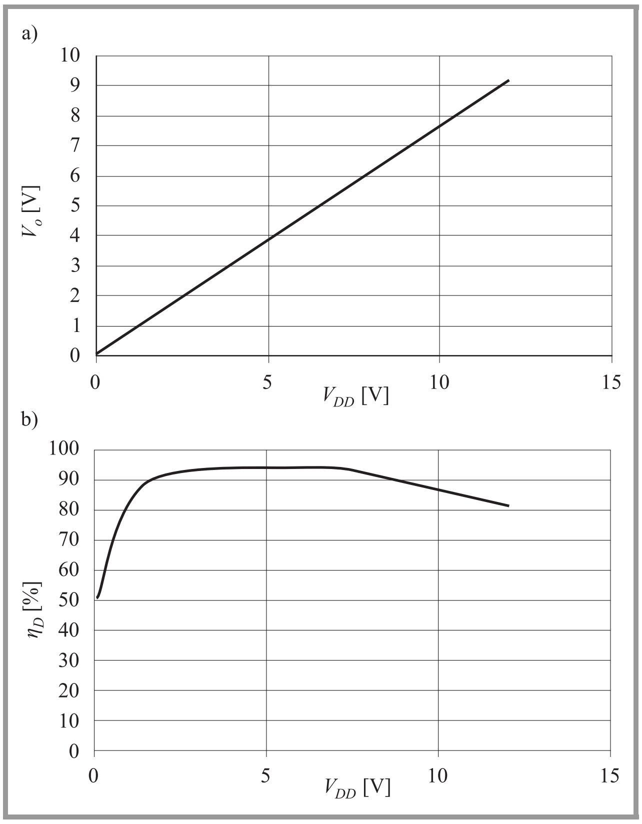


Fig. 10. Characteristic curves (a) $V_o(V_{DD})$ and (b) $\eta_D(V_{DD})$ of the 7 MHz class-E amplifier obtained by PSPICE simulations.

the non-ZVS operation appears already for $V_{DD} \leq 6$ V. It should be noted that in the 7 MHz amplifier for the sufficiently low V_{DD} the transistor turns on in very adverse conditions: at $v_{DS} \gg V_{DD}$ (Fig. 9b). Therefore the efficiency of the 7 MHz class-E amplifier decreases significantly with the supply voltage.

To determine the drain-modulation static characteristic curve $V_o(V_{DD})$ and drain-efficiency static characteristic curve $\eta_D(V_{DD})$ the PSPICE simulations were carried out for $V_{DD} = 0$, 0.1, 0.3 V and from $V_{DD} = 0.5$ V to 12 V every 0.5 V for each amplifier. The obtained results show that the curves $V_o(V_{DD})$ and $\eta_D(V_{DD})$ of the 0.5 MHz class-E amplifier (Fig. 6) are almost perfectly linear with a nearly constant slope: $V_o/V_{DD} \in [0.696,\ 0.714]$ for $V_{DD} \in [0.1\text{ V},\ 12\text{ V}]$, $\eta_D \in [97.44\%,\ 97.90\%]$ for $V_{DD} \in [0.3\text{ V},\ 12\text{ V}]$.

For the 5 MHz amplifier, the characteristic curve $V_o(V_{DD})$ is acceptable: the ratio V_o/V_{DD} increases from 0.681 for $V_{DD} = 1$ V to 0.744 for $V_{DD} = 12$ V (Fig. 8a). Typical nonlinearity can be observed for $V_{DD} < 1$ (similarly as in [3]). Unfortunately, efficiency of the 5 MHz amplifier decreases from about 95% for $V_{DD} > 2$ V to 70.5% for $V_{DD} = 0.1$ V (Fig. 8b).

For the 7 MHz class-E amplifier non-linearity of the characteristic curve $V_o(V_{DD})$ is higher (Fig. 10a) but harmonic distortion of the AM-signal envelope caused by this nonlinearity is acceptable (THD = 1.64% for full drive). In contrast, the efficiency characteristic $\eta_D(V_{DD})$ of the 7 MHz amplifier (Fig. 10b) is not so good, because efficiency of this amplifier decreases sharply with supply voltage (for $V_{DD} < 3$ V). Additionally, its efficiency decrease considerably also for $V_{DD} > 7$ V.

6. Experimental Results

A class-E 12 W/0.5 MHz amplifier (Table 1) was built to verify experimentally the influence of non-linear output capacitance of the transistor on the circuit operation. The measured waveforms of currents and voltages in the amplifier are presented in Fig. 11 for two values of the supply voltage: $V_{DD} = 12$ V and $V_{DD} = 1$ V. It can be noticed that the drain voltage and current waveforms measured for $V_{DD} = 12$ V are in a very good agreement with the PSPICE-simulation results (Fig. 5a). The amplifier operates in the optimum mode and the measured efficiency is high (96%) although it is limited by finite quality factor of the applied coil $L_{sr} = 11.4\ \mu$H with ferrite core.

In contrast, for the very low supply voltage $V_{DD} = 1$ V only the measured waveform of v_{DS} is exactly compatible with the simulation result (Figs. 11b and 5b). The output capacitance C_{oss} of the transistor is large but the total shunt capacitance of the switch increases only a little (Fig. 2, Table 1). Therefore the transistor turns on at non-zero but very low voltage and the measured efficiency of the amplifier is still high (93%). The measured drain-modulation static characteristic curve $V_o(V_{DD})$ of the amplifier is perfectly linear. Unfortunately in the measured waveform of i_D high-level parasitic oscillations are observed (Fig. 11b). These oscillations result from the use of the current probe necessary to measure the drain current. The current probe requires connecting a short wire in series with the drain electrode of the transistor, which effectively introduces a series parasitic inductance in the circuit.

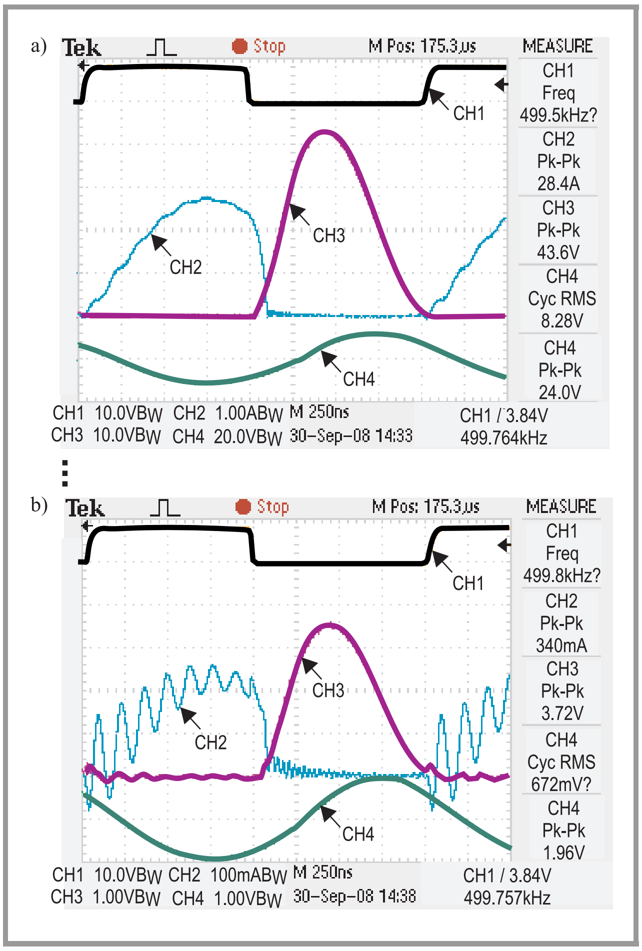


Fig. 11. Waveforms measured in 12 W/0.5 MHz class-E amplifier: (a) $V_{DD} = 12$ V (CH1 – 10 V/div. – v_{GS}, CH2 – 1 A/div. – i_D, CH3 – 10 V/div. – v_{DS}, CH4 – 20 V/div. – v_o); (b) $V_{DD} = 1$ V (CH1 – 10 V/div. – v_{GS}, CH2 – 100 mA/div. – i_D, CH3 – 1 V/div. – v_{DS}, CH4 – 1 V/div. – v_o).

This effect makes difficult the experimental verification for the class-E amplifiers operating at 5 and 7 MHz in which the parasitic inductance caused by the current probe significantly distorts the non-optimum, sub-optimum and even the optimum operation. Therefore the measurement of the class-E amplifiers operating at 5 MHz and 7 MHz will be a subject of separate research. However, the measured characteristics $V_o(V_{DD})$, $\eta(V_{DD})$ of built 5 MHz and 7 MHz class-E amplifiers are very close to the simulation results (Figs. 8 and 10).

7. Conclusions

Simulated results obtained for class-E amplifiers operating at 0.5 MHz, 5 MHz and 7 MHz have shown that all the amplifiers can be applied in power amplitude modulators. The non-linear output capacitance of the transistor in the class-E amplifier does not cause significant distortions of the output-signal envelope in such modulators. However, at high operating frequencies when the transistor non-linear output capacitance becomes a major part of the amplifier shunt capacitor a significant reduction of amplifier power efficiency has been observed for both low and high values of the supply voltage. This phenomenon results from the fact that the range of change of the non-linear transistor output capacitance during operation of the class-E amplifier increases significantly with the supply voltage. Therefore normalized current and voltage waveforms of the class-E amplifier without external shunt capacitance depend on its supply voltage. Thus, the class-E amplifier tuned at a given supply voltage (ensuring ZVS and ZCS turn on) is mistuned for other values of the supply voltage, which is an important issue in HF AM modulators. The problem will be a subject of further research.

References

[1] M. Albulet and S. Radu, "Second order effects in collector amplitude modulation of class E power amplifier", *Int. J. Electron. Commun. (AEU)*, vol. 49, no. 1, pp. 44–49, 1995.

[2] R. E. Zulinski and J. W. Steadman, "Class E power amplifiers and frequency multipliers with finite dc-feed inductance", *IEEE Trans. Circ. Syst.*, vol. CAS-34, no. 9, pp. 1074–1087, 1987.

[3] M. K. Kazimierczuk, "Collector amplitude modulation of the class E tuned power amplifier", *IEEE Trans. Circ. Syst.*, vol. CAS-31, no. 6, pp. 543–549, 1984.

[4] "Power MOSFET catalogue", International Rectifier.

[5] M. K. Kazimerczuk and D. Czarkowski, *Resonant Power Converters*. New York: Wiley, 1995.

[6] M. K. Kazimierczuk and K. Puczko, "Exact analysis of class E tuned power amplifier at any Q and switch duty cycle", *IEEE Trans. Circ. Syst.*, vol. CAS-34, no. 2, pp. 149–159, 1987.

[7] M. J. Chudobiak, "The use of parasitic nonlinear capacitors in class E amplifiers", *IEEE Trans. Circ. Syst. I, Fundam. Theory Appl.*, vol. 41, no. 12, pp. 941–944, 1994.

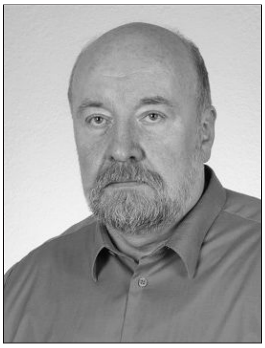
[8] P. Alinikula, K. Choi, and S. Long, "Design of class E power amplifier with nonlinear parasitic output capacitance", *IEEE Trans. Circ. Syst. II, Anal. Dig. Sig. Process.*, vol. 46, no. 2, pp. 114–119, 1999.

[9] T. Suetsugu and M. K. Kazimierczuk, "Comparison of class-E amplifier with nonlinear and linear shunt capacitance", *IEEE Trans. Circ. Syst. I, Fundam. Theory Appl.*, vol. 50, no. 8, pp. 1089–1097, 2003.

[10] T. Suetsugu and M. K. Kazimierczuk, "Analysis and design of class E amplifier with shunt capacitance composed of nonlinear and linear capacitance", *IEEE Trans. Circ. Syst. I, Fundam. Theory Appl.*, vol. 51, no. 7, pp. 1261–1268, 2004.

[11] "Library of power MOSFET models", OrCAD, Inc., 1998.

[12] J. Modzelewski and M. Mikołajewski, "Class-E tuned amplifiers in power amplitude modulators", in *Proc. XVII Int. Conf. Microw. Radar Wirel. Commun. MIKON 2008*, Wrocław, Poland, 2008, vol. 3, pp. 721–724.



Juliusz Modzelewski was born in Warsaw, Poland, in 1953. He received the M.Sc. and Ph.D. degrees in electronics engineering from the Department of Electronics, Warsaw University of Technology in 1977 and 1993, respectively. Since 1977 he has been working in the Institute of Radioelectronics, Warsaw University of Technology (as an Assistant from 1977 to 1993 and as an Assistant Professor from 1993). His teaching, research and development activities are in RF power technology, radio transmitters, and HF high-efficiency power amplifiers.

e-mail: juliuszm@ire.pw.edu.pl
Institute of Radioelectronics
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Mirosław Mikołajewski was born in Warsaw, Poland, in 1961. He received his M.Sc. and Ph.D. degrees from the Warsaw University of Technology in 1987 and 1993, respectively, both in electronics engineering. Since 1988 he has been with Warsaw University of Technology, where he currently holds the position of an Assistant Professor. His field of interest covers resonant HF high-efficiency amplifiers as well as switch-mode DC/DC converters.

e-mail: M.Mikolajewski@ire.pw.edu.pl
Institute of Radioelectronics
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Cooperative and Non-cooperative, Integrative and Distributive Market Games with Antagonistic and Altruistic, Malicious and Kind Ways of Playing

Sylwester Laskowski

Abstract—The article illustrates distinctions between important concepts of game theory, which support understanding the relation between subjects on competitive and regulated telecommunications services market. Especially it shows that often used distinction between retail and wholesale market that treat them respectively as competitive and cooperative can be misleading or even wrong.

Keywords— ***antagonism and altruism, competition, cooperative games, integrative and distributive processes, non-cooperative games.***

1. Introduction

Liberalization of the telecommunications services market transformed so far monopolistic market into competitive one. However it is a specific competition because market players are also forced (by the telecommunications low and decisions of the regulators) into cooperation: networks of the operators ought to be interconnected. For these reasons telecommunications market is not only competitive, but also cooperative.

Cursory analysis leads us to conclusion, that the boundary between the issues of competition and cooperation runs the same way as the boundary between the retail and the wholesale market. However more careful analysis shows that it should not be true. In fact competition is not an opposing part to the cooperation: these concepts comes from different "layers" of interaction between players.

The article discusses three layers, that defines the complexity of interaction between players in market games: possibility of concluding enforceable agreements outside the formal rules of the game, the structure of the payoff matrix, the goals of the players, and explains the essence of the important phenomenon that occurs on each of them.

2. Simple Theoretical Model of a Market Game

Let us describe the market game in the concepts of game theory. Every market participator can be treated as a player, which has his own strategy of playing (e.g., prices on the retail market, interconnection fees on the wholesale market, etc.). The players evaluate their decisions (set strategies) by the single-criteria or aggregated, multiple-criteria goal function, which can be called as their payoff function. The value of the payoff function depend on the strategies set by each player in the game.

Table 1
Relationship between concepts of strategy and outcomes of payoff function

Strategies	b_1	b_2	$\boldsymbol{b_3}$	b_4
a_1			$\vdots$	
$\boldsymbol{a_2}$	......	......	$[\boldsymbol{V_3^A(a_2), V_2^B(b_3)}]$	......
a_3			$\vdots$	
a_4			$\vdots$	

For the case of two players it is useful to illustrate relation between strategies and payoff functions in the form of the so called payoff matrix. Table 1 illustrates a simple payoff matrix for two market players – A and B. Player A chooses one of four strategies: a_1, a_2, a_3, a_4, and player B one of b_1, b_2, b_3, b_4. Choosing the strategy a_i by player A and b_j by player B results in obtaining $V_j^A(a_i)$ by player A and $V_i^B(b_j)$ by player B.

3. Distinction for the sake of the Way of Setting a Solution

From the early work of John Nash [1] differentiation between *cooperative* and *non-cooperative* games starts in game theoretical analysis. In *non-cooperative games* players are unable to conclude enforceable agreements outside the formal rules of the game. *Cooperative games* allow such agreements[1]. Actually, Nash also assumed that in a non-cooperative game, the players will be unable to com-

[1] Nash suggests that non-cooperative games are more basic, that cooperative games may fruitfully be analyzed by reformulating them as non-cooperative ones and by solving for the Nash equilibria [2]. This approach has come to be known as the Nash program [1]. It allows unification of the theory and enables better understanding of the different solution concepts that have been proposed in cooperative theory.

municate with each other. Yet, as it was noticed by Harold W. Kuhn [1], this would be a needlessly restrictive assumption. For if the players cannot enter into enforceable agreements, then their ability to communicate will be of no real help toward a cooperative outcome.

Such differentiation – differentiation on the level of possibility to conclude enforceable agreements – explains us most of all the *way* of setting the result of the game. Such two different ways – with and without agreements – have found reflection into two different theoretical methods of solving games which are now called Nash solution (for cooperative games) [2] and Nash equilibrium (for non-cooperative ones) [3], [4].

Competition is the relation between two or more subjects (players) which arises when such players strives for the same and limited goods. So in terms of game theory such relation is well described by model of zero-sum (or constant-sum) game [2]: the more one player gets the more the other (others) should loose. An example of such game we have in Table 2.

Table 2
Competitive, zero-sum game

Strategies	b_1	b_2
a_1	[–1, 1]	[4, –4]
a_2	[2, –2]	[–3, 3]

So using the above mentioned concepts we should say that popular distinction: *cooperative – wholesale market* and *competitive – retail market* is not too precise. Rather we should say: *cooperative – wholesale market* and *non-cooperative retail market*[2].

4. Distinction for the sake of the Structure of the Payoff Matrix

Another concepts useful for understanding discussed issues comes from the theory of negotiations [4]. Negotiations are in fact *cooperative game* (as on the wholesale telecommunications services market). In every negotiations, where are discussed at least two different issues, and where preferences of the parts are not strictly the same, it is possible to engage the negotiators into a process which is called *integration*. During it parts tries to find such correlation between their preferences which enables *increasing* the size of the "cake" before *dividing* it. In fact such process bases on mutual exchange less preferable issues (or their parts) on more preferable ones. Process of dividing such "cake" (occurring during the integrative process or not) is called *distributive*, and so in fact it is exact *competitive process*.

Saying in terms of game theory or decision support *integrative* process means seeking for effective, Pareto-optimal solutions [5], whilst *distributive* process means making actions for choosing one of two different (and differently preferable by players) solutions (effective or not). It is interesting to notice, that zero-sum game is a game, where every result is Pareto-optimal, and so there exist a place only for *distributive* process.

In negotiations *distributive* process is the only if negotiations concern only one issue [4]. In such situations there is no place for any *integration*, for increasing the size of the "cake". Such "cake" can be only divided and the more one part gets the more the other looses. So every solution, every division is Pareto-optimal. However it is true only if the whole "cake" was divided, if "no gold was left on the table" [6], [7] (strictly speaking only such situations can be modeled as a zero-sum games). If it is possible to exclude some part of the "cake" from the division (from the *distributive* process), then we have place for something like *de-integrative* process which is strictly opposite to the *integrative* one. An example of such game we have in Table 3.

Let us assume, that the values of outcomes corresponds to the share of the divided (distributed) object. In this game there are three effective solutions, that "divide the whole cake": [0.3, 0.7], [0.7, 0.3] and [0.5, 0.5]. The result [0.4, 0.4] is not effective, and means "leaving on the table some gold" (0.2 part of the object), and so choosing it (if players does know the effective results) can be treated as a result of a *de-integrative* process.

Table 3
Game with possibility of de-integration

Strategies	b_1	b_2
a_1	[0.4, 0.4]	[0.5, 0.5]
a_2	[0.3, 0.7]	[0.7, 0.3]

Is there any opposite process to the *distribution* in such games? Generally we say that this is *concentration*, but in a game it means no decision, giving up any solution, and so, from our point of view it is not an interesting case. However as we have games where there is only place for *distributive* process (zero-sum games), so we have games where there is no place for any *distribution*, only an *integration* (and *de-integration*) can take place. An example of such game we have in Table 4, where – assuming that both players aim only at maximizing their own payoffs – preferences of both players between individual solutions are exactly the same. If we assume that players aims at choosing an effective result, and are able (during the *integrative* process) to doing so, there is no place for *distributive* process, because there

Table 4
Game with no place for distributive process

Strategies	b_1	b_2
a_1	[1, 1]	[4, 4]
a_2	[2, 2]	[3, 3]

[2] If really decisions on the retail market are made by players without making any (public or tacit) agreements.

is only one effective result in this game. More over making a comparison between any two different solutions we have that always one of them is betters for both players.

In some cases the place for *distributive* process can arise if *integrative* process finishes without finding the only effective solution. An example of such game we have in Table 5.

Table 5
Game with a place for distributive process only if integrative process finishes without finding the only effective result

Strategies	b_1	b_2
a_1	[1, 1]	[4, 4]
a_2	[3, 2]	[2, 3]

If for example during a negotiations, assuming that players aims at maximizing their own payoffs players didn't find (during an *integrative* process) a solution [4, 4], than a place for *distributive* process would arise: players would bargain on selections one of two, differently preferable solutions: [3, 2] and [2, 3].

Analogically it is also possible that there would be no place for a *distributive* process because of the same reason. We have this in a game as in Table 6.

Table 6
Game without a place for distributive process if players didn't find all effective solutions

Strategies	b_1	b_2
a_1	[1, 1]	[2, 2]
a_2	[3, 4]	[4, 3]

If, for example during the negotiations players found only two or three results (but only one effective: [3, 4] or [4, 3]), then there would be no place for a *distributive* process.

It seems to be useful to define a single type of integrative-distributive games, encompassing three before mentioned cases:

- game with possibility of de-integration,
- game with a place for distribution only if integrative process finishes without finding the only effective result,
- game without a place for distribution if players didn't find all effective results.

In fact an *integrative* process can occur if there exists a place for improving a given solution for every of the players at a given stage of a game. Starting from any ineffective result always such possibility exists. In every above mentioned cases there exists at least one ineffective result, so such situation occurs. Also in every cases there are at least two incomparable solutions – solutions, that any of them is better for every of the players – so there exists a place for a *distributive process* (it's true even if there exists only one effective result in a game, because players should not know, which result is effective, an can decide to finish integrative process after finding two different and incomparable ineffective results).

So, respectively to the structure of the payoff matrix of a given game we can distinguish the following different types of games:

- *distributive games*: games with no place for *integrative* process (strictly competitive games, zero-sum or constant-sum games),
- *integrative games*: games with no place for *distributive* process (not competitive games),
- *integrative-distributive games*.

Having this we can say that *wholesale* telecommunications services market, respectively to structure of the payoff matrix can be treated as *distributive*, *integrative* or *integrative-distributive cooperative game*. Analogically *retail* telecommunications services market, can be treated as *distributive*, *integrative* or *integrative-distributive non-cooperative game*. So we see that as on the retail (*non-cooperative*) so on the wholesale (*cooperative*) market a place for a *distributive* process – a real *competition* – can exist. Also we see, that on the retail (*non-cooperative*) market there can exist a place for something like *integration*, for increasing the size of the "cake", for finding such solution which would be better for every of the players than another accessible solution.

However there is a difference between *integrative* and *distributive* processes on retail and wholesale markets. Such difference comes from the *way* of setting a result: *cooperative* on the wholesale markets and *non-cooperative* on the retail markets. On the wholesale markets *integrative* and *distributive* process can proceed during one, single game (during a one round of the negotiations by making temporary decisions). On the retail markets such process proceeds only if a game is repeated (by making real decisions). Moreover, on the wholesale markets *integrative* and *distributive* processes can proceed independently. On the retail markets such *integration* and *distribution* are realized simultaneously: by making a decision by the last mover in the game[3].

5. Distinction for the sake of the Aims of the Players

Until to the first works of John C. Harsanyi on the games with incomplete information (so called *I-games* [8]–[11]) it was generally assumed, that in any games players have all information, necessary to define the strategic form of a given game (its *basic mathematical structure*).

[3] Partially *integration* and *distribution* are realized also by a decisions of the first (and eventually next, but not last) mover, whose decisions creates the finale alternatives to the last mover.

Harsanyi showed that in many real situations this is too hard assumption. Players often does not know: the form of their own or the other player's payoff function, the set of the accessible strategies, the scope of information that the other player possess, etc.

In real situations there exists one thing, that really can be interpreted as a part of a strategic form of a game, yet seems to be simple to pass over. General assumption in game theory is that players aims at maximizing their own payoff functions. These functions – interpreted as utility functions – are formulated in such way, that their maximization leads to obtaining the appointed goal. Such utility function describes how good for a given player (under his subjective preferences) is the obtained objective state. Here arises very subtle problem.

Let us consider simple example. In a given game there are only two different results [3, 4] and [1, 3]. The values reflects the profits in money of players A and B. Let us assume, that both players prefer to get more money than less, and that their utility is proportional to the amount of gotten many. So such results expressed in terms of utility have the same form: [3, 4] and [1, 3]. The answer, which result should be chosen by players seems to by simple: effective [3, 4]. However it is true only, if – as it is usually assumed in game theory analysis, and us during formulation of utility function for these results – players evaluate the results only by the value of money, they get themselves.

This assumption can be called as assumption of *neutral* way of playing, by players: players are interesting only in evaluation the values gotten by themselves.

In market games such assumption is too hard. Evaluation solely the values gotten by itself is a good approach only in short term. In long term players should take into consideration relative values, because after crossing a certain distance between the positions of the players on the market, such distance can increase very quickly: strong player becomes stronger, weak becomes weaker. So in our example we could assume, that players evaluate the obtained results not by the values of money obtained by themselves but as a difference between the values gotten by both players. So for player A the result [3, 4] may have utility $3-4=-1$, and for result [1, 3]: $1-3=-2$. For player B the utilities would be exactly opposite: for [3, 4]: $4-3=1$, and for [1, 3]: $3-1=2$. These new utility function defines in fact different solutions (in terms of utility): for the values [3, 4] now the result in terms of utility is $[-1, 1]$ and for [1, 3] – $[-2, 2]$. Both of them are effective.

Such aspirations of players can be called as *antagonistic*. Generally, when we say antagonism of the player, we mean of the situation, when the player aims not only in maximization of his own payoff function (defined as an evaluation only his own vale), but also in minimization of the other player's payoff function (defined in the same way). As an opposition we can formulate an aspiration which can be called as *altruistic*. In such a case a given player would aspire to maximize the payoff function of the other player. Now we formulate in mathematical form some examples of the *antagonistic* and *altruistic* way of playing, which can we called antagonistic and altruistic aims. For the simplicity we formulate them only for the player B.

5.1. Examples of Antagonistic Aims

Antagonistic aim of player B reflects his approach to his own payoff function and to player's A payoff function. There could be many of such aims. Below we will present some of them.

Let's $\breve{b}_k$ be the (kth) antagonistic strategy (move) of player B. The most antagonistic move of player B is such, that B aims first of all at minimization of the A's payoff function, and he considers his own payoff function only in a case of ambiguity (two or more different strategies give the same and the smallest outcome to player A). This can be expressed as follows:

$$\breve{b}_k(a_i) = \arg\operatorname{lex}\min_j \left\{ V_j^A(a_i), -V_i^B(b_j) \right\}. \tag{1}$$

The least antagonistic move of player B is such, that B aims first of all at maximization of his own payoff function and in the case of ambiguity (two or more different strategies give the same – and the highest – outcome to him) he chooses this, that gives the smallest outcome to player A. This can be expressed as follows:

$$\breve{b}_k(a_i) = \arg\operatorname{lex}\max_j \left\{ V_i^B(b_j), -V_j^A(a_i) \right\}. \tag{2}$$

Strategies (1) and (2) determine (for a given strategy of player A) the range of outcomes that player A can obtain in a situation that player B plays in an antagonistic way. Below some other antagonistic aims of player B are described.

Player B can aim at maximizing of his own payoff function and at minimizing of player's B payoff function with different power to both of them expressed by a weight coefficient α. In such a way a general form of a formula (1) can be obtained[4]:

$$\breve{b}_k(a_i) = \arg\max_j \left\{ \alpha \cdot V_i^B(b_j) - (1-\alpha) \cdot V_j^A(a_i) \right\}. \tag{3}$$

Strategy (3) can be interpreted as aiming at maximizing the difference between the outcomes of player B and A.

Player B can also aim at obtaining assumed value of the difference – δ between the outcomes of the players, and after that at maximizing of his own payoff function. This can be expressed as the following lexicographic optimization task:

$$\breve{b}_k(a_i) = \arg\operatorname{lex}\max_j \left\{ \Delta_{ij}, V_i^B(b_j) \right\}, \tag{4}$$

where:

$$\Delta_{ij} = \min \left\{ \delta, \alpha \cdot V_i^B(b_j) - (1-\alpha) \cdot V_j^A(a_i) \right\}.$$

Another kind of antagonistic strategy can be expressed as aiming at maximization of an own payoff function with simultaneous aiming at ensuring that the other player's payoff

[4] The formula (1) can be generalized to (3) by assumption $\alpha \gg (1-\alpha)$.

function does not exceed assumed threshold value ν. This can be expressed as the following optimization task:

$$\breve{b}_k(a_i) = \arg\max_j \left\{ V_i^B(b_j) \right\}, \tag{5}$$

under constraint:

$$V_j^A(a_i) \leq \nu.$$

There is a possibility to make an opposite approach: minimization of the player's A payoff function, under assumption that the outcome of player B would not be smaller then the threshold value ν:

$$\breve{b}_k(a_i) = \arg\min_j \left\{ V_j^A(a_i) \right\}, \tag{6}$$

under constraint:

$$V_i^B(b_j) \geq \nu.$$

In the case of using strategy (5) or (6) it is important to asses correctly the value of the threshold ν in order to assure that the appropriate optimization problems will have a solution.

It is possible to express the antagonistic approach of the player B with using the concepts of reference point method [5], [12] by introducing reservation and aspiration point for the payoff functions of the player A and B. Payoff function of the player A will be treated here as the minimized criterion and the player's B as the maximized criterion. Partial achievement function for player B is then expressed as follows:

$$\eta_B\left(V_i^B(b_j)\right) = \begin{cases} \frac{\beta(V_i^B(b_j) - \underline{V}^B)}{\overline{V}^B - \underline{V}^B} & \text{for} \quad V_i^B(b_j) < \underline{V}^B \\ \frac{V_i^B(b_j) - \underline{V}^B}{\overline{V}^B - \underline{V}^B} & \text{for} \quad \underline{V}^B \leq V_i^B(b_j) \leq \overline{V}^B \\ 1 + \frac{\alpha(V_i^B(b_j) - \overline{V}^B)}{\overline{V}^B - \underline{V}^B} & \text{for} \quad \overline{V}^B < V_i^B(b_j), \end{cases} \tag{7}$$

where $\underline{V}^B$ represents reservation point, and $\overline{V}^B$ represents aspiration point for the payoff function $V_i^B(b_j)$ of player B.

Partial achievement function for player A is expressed as

$$\eta_A\left(V_j^A(a_i)\right) = \begin{cases} 1 + \frac{\alpha(V_j^A(a_i) - \overline{V}^A)}{\overline{V}^A - \underline{V}^A} & \text{for} \quad V_j^A(a_i) < \overline{V}^A \\ \frac{V_j^A(a_i) - \underline{V}^A}{\overline{V}^A - \underline{V}^A} & \text{for} \quad \overline{V}^A \leq V_j^A(a_i) \leq \underline{V}^A \\ \frac{\beta(V_j^A(a_i) - \underline{V}^A)}{\overline{V}^A - \underline{V}^A} & \text{for} \quad \underline{V}^A < V_j^A(a_i). \end{cases} \tag{8}$$

In such a case antagonistic response (antagonistic strategy) of player B can be defined as the following formulae:

$$\begin{aligned} \breve{b}_k(a_i) = \quad & \arg\max_j \Bigg\{ \min\left\{ \eta_A\left(V_j^A(a_i)\right), \eta_B\left(V_i^B(b_j)\right) \right\} \\ & + \rho \cdot \left(\eta_A\left(V_j^A(a_i)\right) + \eta_B\left(V_i^B(b_j)\right) \right) \Bigg\}. \end{aligned} \tag{9}$$

5.2. Examples of Altruistic Aims

Let's $\widehat{b}_k$ be the (kth) altruistic strategy (move) of player B. The most altruistic move of player B is such, that B aims first of all at maximization of the A's payoff function, and he considers his own payoff function only in a case of ambiguity (two or more different strategies give the same and the highest outcome to player A). This can be expressed as

$$\widehat{b}_k(a_i) = \arg\operatorname{lex}\max_j \left\{ V_j^A(a_i), V_i^B(b_j) \right\}. \tag{10}$$

The least altruistic move of player B is such, that B aims first of all at maximization of his own payoff function and in the case of ambiguity (two or more different strategies give the same – and the highest – outcome to him) he chooses this, that gives the highest outcome to player A. This can be expressed as follows:

$$\widehat{b}_k(a_i) = \arg\operatorname{lex}\max_j \left\{ V_i^B(b_j), V_j^A(a_i) \right\}. \tag{11}$$

Strategies (10) and (11) determine (for a given strategy of player A) the range of outcomes that player A can obtain in a situation that player B plays in an altruistic way. Below some other altruistic moves of player B are described.

Player B can aim at maximizing of his own payoff function and at maximizing of player's B payoff function with different power to both of them expressed by a weight coefficient α. In such a way a general form of a formula (10) can be obtained[5]:

$$\widehat{b}_k(a_i) = \arg\max_j \left\{ \alpha \cdot V_i^B(b_j) + (1 - \alpha) \cdot V_j^A(a_i) \right\}. \tag{12}$$

Strategy (12) can be interpreted as aiming at maximizing the sum of the outcomes of player B and A.

Another kind of altruistic strategy can be expressed as aiming at maximization of an own payoff function with simultaneous aiming at ensuring that the other player's payoff function will be not smaller than the assumed threshold value ν. This can be expressed as the following optimization task:

$$\widehat{b}_k(a_i) = \arg\max_j \left\{ V_i^B(b_j) \right\}, \tag{13}$$

under constraint:

$$V_j^A(a_i) \geq \nu.$$

There is a possibility to make an opposite approach: maximization of the player's A payoff function, under assuming that the outcome of player B would not be smaller then the threshold value ν:

$$\widehat{b}_k(a_i) = \arg\max_j \left\{ V_j^A(a_i) \right\}, \tag{14}$$

under constraint:

$$V_i^B(b_j) \geq \nu.$$

[5] The formula (10) can be generalized to (12) by assumption $\alpha \gg (1 – \alpha)$.

In the case of using strategy (13) or (14) it is important to correctly asses the value of the threshold ν in order to assure that the appropriate optimization problems will have a solution.

It is possible to express the altruistic approach of the player B with using the concepts of reference point method by introducing reservation and aspiration point for the payoff functions of the player A and B. Payoff functions of the player A and B are treated here as the maximized criterion. Partial achievement function for player B is then expressed as follows:

$$\eta_B\left(V_i^B(b_j)\right)=\begin{cases}\frac{\beta(V_i^B(b_j)-\underline{V}^B)}{\overline{V}^B-\underline{V}^B} & \text{for} \quad V_i^B(b_j)<\underline{V}^B\\ \frac{V_i^B(b_j)-\underline{V}^B}{\overline{V}^B-\underline{V}^B} & \text{for} \quad \underline{V}^B\leq V_i^B(b_j)\leq\overline{V}^B\\ 1+\frac{\alpha(V_i^B(b_j)-\overline{V}^B)}{\overline{V}^B-\underline{V}^B} & \text{for} \quad \overline{V}^B<V_i^B(b_j),\end{cases} \tag{15}$$

where $\underline{V}^B$ represents reservation point, and $\overline{V}^B$ represents aspiration point for the payoff function $V_i^B(b_j)$ of player B.

Partial achievement function for player A is expressed as

$$\eta_A\left(V_j^A(a_i)\right)=\begin{cases}\frac{\beta(V_j^A(a_i)-\underline{V}^A)}{\overline{V}^A-\underline{V}^A} & \text{for} \quad V_j^A(a_i)<\underline{V}^A\\ \frac{V_j^A(a_i)-\underline{V}^A}{\overline{V}^A-\underline{V}^A} & \text{for} \quad \underline{V}^A\leq V_j^A(a_i)\leq\overline{V}^A\\ 1+\frac{\alpha(V_j^A(a_i)-\overline{V}^A)}{\overline{V}^A-\underline{V}^A} & \text{for} \quad \overline{V}^A<V_j^A(a_i).\end{cases} \tag{16}$$

In such a case altruistic move of player B can be defined as the following formulae:

$$\begin{aligned}\widehat{b}_k(a_i) \quad = \quad & \arg\max_j\left\{\min\left\{\eta_A\left(V_j^A(a_i)\right),\eta_B\left(V_i^B(b_j)\right)\right\}\right.\\ & \left.+\rho\cdot\left(\eta_A\left(V_j^A(a_i)\right)+\eta_B\left(V_i^B(b_j)\right)\right)\right\}.\end{aligned} \tag{17}$$

5.3. Examples of Irrational Aims

As an irrational way of playing we mean such, that a given player aims most of all at minimizing his own payoff function. It should be stressed that as in antagonistic so in altruistic ways of playing there is a place for deteriorating of the own payoff. However it is rather a consequence of the main goal: decreasing (in antagonistic) or increasing (in altruistic) the payoff of the other player. If such deteriorating of the own payoff couldn't find any justification in such mainly antagonistic or altruistic aims, than we should treat it as irrational.

Here we present some examples of irrational aims:

$$\tilde{b}_k(a_i)=\arg\operatorname{lex}\min_j\left\{V_i^B(b_j),V_j^A(a_i)\right\}, \tag{18}$$

$$\tilde{b}_k(a_i)=\arg\operatorname{lex}\min_j\left\{V_j^A(a_i),V_i^B(b_j)\right\}, \tag{19}$$

$$\tilde{b}_k(a_i)=\arg\operatorname{lex}\max_j\left\{V_j^A(a_i),-V_i^B(b_j)\right\}, \tag{20}$$

$$\tilde{b}_k(a_i)=\arg\operatorname{lex}\max_j\left\{-V_i^B(b_j),V_j^A(a_i)\right\}, \tag{21}$$

$$\tilde{b}_k(a_i)=\arg\min_j\left\{V_j^A(a_i)\right\}, \tag{22}$$

under constraint:

$$V_i^B(b_j)\leq\nu,$$

$$\tilde{b}_k(a_i)=\arg\max_j\left\{V_j^A(a_i)\right\}, \tag{23}$$

under constraint:

$$V_i^B(b_j)\leq\nu.$$

5.4. Context Relative Aim

Let us look once again on two before defined strategies:

$$\breve{b}_k(a_i)=\arg\max_j\left\{\alpha\cdot V_i^B(b_j)-(1-\alpha)\cdot V_j^A(a_i)\right\} \tag{24}$$

and

$$\widehat{b}_k(a_i)=\arg\max_j\left\{\alpha\cdot V_i^B(b_j)+(1-\alpha)\cdot V_j^A(a_i)\right\}. \tag{25}$$

It was said that antagonistic strategy (24) can be interpreted as aiming at maximizing the difference between the outcomes of player B and A. Altruistic strategy (25) can be interpreted as aiming at maximizing the sum of the outcomes of player B and A.

Looking on these we can simply formulate a strategy which in fact can't be unambiguously classified as antagonistic or altruistic, and which is not a irrational one. We mean of a strategy defined as minimization of the difference between payoffs:

$$b_k(a_i)=\arg\min_j\left\{\alpha\cdot V_i^B(b_j)-(1-\alpha)\cdot V_j^A(a_i)\right\}. \tag{26}$$

Minimization of the difference between payoffs seems to be an example of the altruistic strategy: a given player (B) is willing to decrease his own outcome and at the same time to increase the outcome of the other player, in order to ensure the smallest difference between them. However

Table 7
A game where minimization of the difference between the outcome of player B and A can't be interpreted as an altruistic aim

Strategies	b_1	b_2
a_1	[1, 1]	[2, 3]

in some cases such aim gives a solution which should be interpreted as a result of antagonistic or irrational aim. Let us consider a game with payoff matrix like in Table 7.

For the simplicity we assumed that there is only one strategy of player A. Aim defined by strategy (26) leads to the result: [1, 1], which minimizes the difference between the payoffs. However this solution can't be interpreted as a result of altruism of the player B. In fact the payoff of player A is worse than it would be for a [2, 3]. If however a payoff matrix would be like in Table 8, then aim defined by strategy (26) leads to [3, 3], which can be interpreted as a result of altruistic move of player B.

Table 8
A game where minimization of the difference between the outcome of player B and A can be interpreted as an altruistic aim

Strategies	b_1	b_2
a_1	[3, 3]	[2, 3]

So we see, that interpretation of strategy (26) depends on the form of the payoff matrix, and so is *context relative*. This can also lead to misleading the real motives of choosing given strategies by players.

5.5. Outside the Mathematical Structure of a Game: Malicious and Kind Aims

Now we ask an important question: are aims of the players (*antagonistic* or *altruistic*) a part of a *basic mathematical structure* of a game or are they outside it? Or in different way: can we transform games with such aims of players into games with new payoff function and *neutral* aims of the players? The answer seems to be ambiguous.

From one point of view we can say like that: the aim of a player can be simply expressed in values of a utility. In fact the above mentioned *antagonistic* and *altruistic* aims have defined real utility of any solutions, described in terms of utility with *neutral* aim.

Let us remind the before considered example. In a given game there are only two different results: [3, 4] and [1, 3]. The values reflect the profits in money of players A and B. Both players prefer to get more money than less, and their utility is proportional to the amount of gotten money (*neutral aim*). So such results expressed in terms of utility have the same form: [3, 4] and [1, 3]. However if for example player A aims into an *antagonistic* aim defined as aiming at maximization the difference between the outcomes of the players, than in fact he has different utility function (payoff function), defined as a difference between the utility values expressed with assumption of a *neutral* aim. So we can incorporate an *antagonistic* aim of player A into his payoff function, and treat this new situation as a game with *neutral* aims of the players. In such a game the solutions will be expressed in form $[3-4,\,4] = [-1,\,4]$ and $[1-3,\,3] = [-2,\,3]$. So we see that different than *neutral* aims of the players can be simply incorporated into a payoff function of a player and so can be treated as a part of a *basic mathematical structure* of a game.

However there are two aims of the players, which may cause a problem with incorporation them into a payoff function of a players, and so which seems to be outside the *basic mathematical structure* of a game. We call them: *malicious* and *kind* way of playing.

A *malicious* way of playing means that a given player *defines* his own aim as the opposite of the aim of the other player. A *kind* way of playing means that a given player *defines* his own aim as an exact realization of the other's player aim. Of special importance is here the word *defines*. Such word justifies why such way of playing are not called merely as *opposing* and *convergent*. For example, if both players are going to play in the least antagonistic way, and so aims at choosing the following strategies:

$$\breve{b}_k(a_i) = \arg\operatorname{lex}\max_{j}\left\{V_i^B(b_j), -V_j^A(a_i)\right\}, \tag{27}$$

$$\breve{a}_k(b_j) = \arg\operatorname{lex}\max_{i}\left\{V_j^A(a_i), -V_i^B(b_j)\right\}, \tag{28}$$

the aims:

$$\text{aim}_B = \operatorname{lex}\max_{j}\left\{V_i^B(b_j), -V_j^A(a_i)\right\},$$

$$\text{aim}_A = \operatorname{lex}\max_{i}\left\{V_j^A(a_i), -V_i^B(b_j)\right\}$$

are really *opposing*, but we can say that (for example) player A plays in a *malicious* way if he explicitly defines his aim as

$$\text{aim}_A = \sim \text{aim}_B\,,$$

where $\sim$ means *opposing of* (independently of in which way aim_B would be defined). Analogically we could say that player A plays in a *kind* way only if he explicitly defines his aim as

$$\text{aim}_A = \text{aim}_B\,.$$

Malicious and *kind* aims can be also incorporated into payoff functions of players. If for example player A aims at a *neutral* aim defined as maximizing of his own payoff function V^A, then a *malicious* aim of player B can be incorporated into his payoff function by putting $V^B = -V^A$, and treated this function as a function in game with a *neutral aim*. Analogically we can express a *kind* aim of player B by putting $V^B = V^A$. So we see that games with *malicious* aims can be treated as a *distributive* games (zero-sum, strictly competitive) with *neutral* aims, and games with *kind* aims as an *integrative* game (with no place for any *distribution*).

However a problem arises when both players would like to play in a *malicious* (or *kind*) way. How to define their aims in terms of the value of the payoff function? What is a *basic mathematical structure* of such a game, under assumption that players play in a *neutral* way? We cant find a satisfying answer on these questions.

6. Opposition and Convergence of Aims and Problems with Cooperation

Our earlier analysis appointed three important aspects that define the relation between players, and in fact define the real game:

1. The way of setting a final solution, which can be:
   - *cooperative*: players are able to conclude enforceable agreements outside the formal rules of the game,
   - *non-cooperative*: players are unable to conclude enforceable agreements outside the formal rules of the game.
2. The structure of the payoff matrix, that define the game as:
   - *distributive*: game with no place for integrative process (strictly competitive game, zero-sum or constant-sum game),
   - *integrative*: game with no place for distributive process (not competitive game),
   - *integrative-distributive*: there exists a place as for integration so for distribution.
3. The aims of the players, which can be:
   - *neutral*: a given player is interesting only in his own payoffs and aims at maximizing it,
   - *antagonistic*: a given player aims at minimizing the payoff function of the other player,
   - *altruistic*: a given player aims at maximizing the payoff function of the other player,
   - *irrational*: a given player aims at minimizing of his own payoff function,
   - *context-relative*: a given player aims at minimization of the difference between payoffs of the players,
   - *malicious*: a given player tries to thwart another player's plans,
   - *kind*: a given player tries to help in realizing another player's plans.

Now we will analyze the relation between such aspects.

Players decide to play or are forced (more precisely: should be forced) into playing in *cooperative* way only if it leads to more effective or more fair solution than gotten during a *non-cooperative* playing. Increasing effectiveness of the solution comes from *integrative* process. Increasing fairness of the solution is related with the *distributive* process. *Integration* as a process of increasing the "size of a cake" is profitable for both players. *Distribution*, as a process of dividing such "cake" always mean that the more one player gets the more the other should loose. So increasing the fairness of the solution always mean that during it one player will get more and the other will lose. So we can say, that *cooperation* would be always more simple and more natural in *integrative* games than in *distributive* games. In a *integrative-distributive* games *cooperation* will be desirable by a given player only if he hoped that he got more during an *integrative* process than he could probably lose during a *distributive* one. Of course a player will desire a *cooperation* if he hoped that he increases his payoffs also during a *cooperative-distributive* process.

So we can say, the more *integrative* structure of the payoff matrix the simpler *cooperation* between players. Analogically the more *distributive* structure of the payoff matrix the more difficult *cooperation*, the strongest incentive for one of the players to play in a *non-cooperative* way.

Interesting relation occurs between the aims of the players and the process of *cooperation*. Generally we can say: the more *convergent* aims of the players the simpler *cooperation*, the more *opposing* aims – the more difficult *cooperation*. So we should ask: which aims are convergent, and which are opposing?

It is obvious that if one player played in a *malicious* way then real *cooperation* would be impossible (one player would like to get exactly opposing solution then the other). If one player played in a *kind* way then *cooperation* would be very simple (both players would like to get exactly the same solution). Paradoxically when both players played in *kind* way then *cooperation* may be difficult because of problems with definition of real aim.

It is interesting that also *antagonistic* aims of the players can make *cooperation* simple. In fact *antagonistic* move of one player can be the most desirable from the other player point of view, so antagonistic aim can be treated not as an *opposing* but *convergent*.

Let us consider the following example.

The pay off matrix is like in Table 9.

Table 9
Convergent antagonistic aims

Strategies	b_1	b_2
a_1	[1, 3]	[3, 4]
a_2	[2, 4]	[4, 5]

Let players aim at realizing the following antagonistic goals:

- Player A: maximization of his own payoff function under constraint that the payoff of the player B will be not higher than 4.
- Player B: maximization of his own payoff function under constraint that the difference between the payoffs of the players will be not smaller than 2 (with advantage of the payoff of player A).

Under such assumption, both players would like to set a solution [2, 4]. Under such aims of the players it is the best result in this game for both players, so *cooperation* in this case would be very simple. Obviously *cooperation* would be also simple if both players played in neutral or altruistic way. In such cases both players would like to set a solution [4, 5]. However if player A played in antagonistic way and player B in neutral or altruistic, then *cooperation* could be difficult, because both players would like to set different solution.

This example shows also that in the case of antagonistic aims of both players we have something like *changing the meaning of effectiveness*: both players prefer [2, 4] over [4, 5].

The above discussed example shows us, that in some cases antagonism of the players can make *cooperation* simpler.

It is interesting, that in some cases cooperation with antagonistic aims of both players can be simpler even than in the case of altruistic aims of both of them (not only one of them). Let us consider the following example. The payoff matrix in a game is like in Table 10.

Table 10
Difficult cooperation in the case of altruism of the players

Strategies	b_1	b_2
a_1	[6, 5]	[2, 5]
a_2	[2, 4]	[5, 6]

The players can aim at antagonistic goals:

- Player A: maximization of his own payoff function under constraint that the payoff of the player B will be not higher than 4.
- Player B: maximization of his own payoff function under constraint that the difference between the payoffs of the players will be not smaller than 2 (with advantage of the payoff of player B).

If they are not afraid to disclose them, then they simply find a solution [2, 4], as a satisfying one.

However if the players aims at altruistic aims defined as maximizing the own payoff function under assumption that the other player's payoff would be not smaller than 5, then they would have a problem, which solution should be chose. Player A prefers [6, 5] and player B prefers [5, 6], and if they would not change theirs aims the negotiations may be very strong. In fact above defined antagonistic aims (in this game) were here more convergent than such altruistic aims. So we see that in some cases cooperation among antagonistic players may be simpler than between altruistic ones.

Our conclusion can be justified also in different way. If we transform a game with antagonistic or altruistic aims into a game with new payoff function (which reflects such aims) and neutral aims, then we find that cooperation was simple there where was only one effective solution (where there was no place for distribution). Analogically cooperation was difficult where in such new game (with neutral aims) there was more then one effective result.

7. Summary and Final Conclusions

As it was said in the introduction, cursory analysis of the telecommunications services market leads to conclusion, that the boundary between the issues of *competition* and *cooperation* runs the same way as the boundary between the retail and the wholesale market. Now we see that competition is not an opposing part to the cooperation: in fact these concepts comes from different "layers" of interaction between players. The concept of *cooperation* explains *the way of setting a final result*, while *competition* is in fact a *distribution* process and its existence depends on the *structure of the payoff matrix* and the *aims* of the players. So it is possible, that under some kinds of payoff functions and aims of the players real *competition* can take place as on the retail so on the wholesale markets. It is also possible (even if only theoretically), that competition take place *only* on (*cooperative*) wholesale market, because on (*non-cooperative*) retail market the structure of the payoff matrix and the aims of the players may make a place only for *integrative* process.

On the wholesale telecommunications services market *cooperation* – negotiations on the conditions of the interconnection – is necessary (network ought to be interconnected) and due to unequal distributed negotiations power – forced by the regulator. Our analysis of convergence and opposition of aims of the players shows that such cooperation may be – respectively to the form of payoff matrix and aims of the players – more or less "natural", simple to introducing. Intervention of regulator often stops on the level of *the way of setting a final solution*: players may and ought to negotiate a final solution. Sometimes such intervention changes also the form of the payoff function (e.g., by changing the structure of the cost function, or setting a limitations on the prices). However probably newer such intervention changes the aims of the players. So finally as course of *cooperation* so final result of a game stays difficult to predict.

Probably the most unexpected conclusion of our analysis is that in some cases cooperation between two players which aims at antagonistic goals may be simpler then between players which would like to play in an altruistic way. In fact, under our definition antagonism does not should mean malice of the players – though it might mean. Paradoxically, it is possible that some altruistic aims may express the most malicious way of playing.

Generally as *antagonism* so *altruism* mean that the main player's aim is not objective, but relative: the player evaluates obtained outcome of his payoff function not as an independent single criterion but in comparison to the other player's payoff function (double criteria of evaluation). In this sens, a player which aims at realizing of an antagonistic or altruistic goal doesn't have to know the other player's aim (possibly also antagonistic or altruistic), what should be necessary if he really aims at realizing malice (trying to thwart another player's plans) or kind (trying to help in realizing another player's plans) objective.

References

[1] J. F. Nash, "The work of John Nash in game theory", Nobel Sem., edited version of a seminar devoted to the contributions to game theory of John Nash, Dec. 1994.

[2] P. D. Straffin, *Teoria gier*. Warsaw: Wydawnictwo Naukowe Scholar, 2001 (transl. – in Polish).

[3] J. Lemaire, "Cooperative game theory and its insurance applications", *Astin Bull.*, vol. 21, no. 1, 1991.

[4] H. Raiffa, *The Art and Since of Negotiation*. Cambridge: Harvard University Press, 1982.

[5] A. P. Wierzbicki, "Reference point methods in vector optimization and decision support", Interim Rep. IR-98-017, IIASA, Laxenburg, 1998.

[6] D. Cray and G. E. Kersten, "Negotiating inefficient compromises: is less better than more", Interim Rep. IR-99-022, IIASA, Laxenburg, 1999.

[7] G. E. Kersten and G. R. Mallory, "Rational inefficient compromises in negotiation", Interim Rep. IR-98-024, IIASA, Laxenburg, 1998.

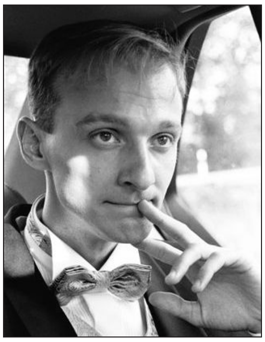
[8] J. C. Harsanyi, "Bargaining in ignorance of the opponent's utility function", *J. Confl. Resol.*, vol. 6, no. 1, pp. 29–38, 1962 [Online]. Available: http://cowles.econ.yale.edu/P/cp/p01b/p0173.pdf

[9] J. C. Harsanyi, "Games with incomplete information – Nobel lecture", 1994 [Online]. Available: http://nobelprize.org/nobel_prizes/economics/laureates/1994/harsanyi-lecture.pdf

[10] R. B. Myerson, "Leartning game theory from John Harsanyi. Notes for a memorial service at Berkeley", Kellogg School of Management, Aug. 2000 [Online]. Available: http://www.kellogg.northwestern.edu/faculty/myerson/research/hars.pdf

[11] R. B. Myerson, "Harsanyi's games with incomplete information", *Manage. Sci.*, no. 50, pp. 1818–1824, 2004 [Online]. Available: http://home.uchicago.edu/~rmyerson/research/harsinfo.pdf

[12] A. P. Wierzbicki and M. Makowski, "Multi-objective optimization in negotiation support", Tech. Rep., IIASA, Laxenburg, 1992.



Sylwester Laskowski received his M.Sc. (1999) in telecommunications from the Institute of Telecommunications of Warsaw University of Technology, Poland, M.A. (2003) in classical guitar from the Warsaw Academy of Music, and Ph.D. (2006) in telecommunications from the Institute of Control and Computation Engineering of Warsaw University of Technology. His current research interests concentrate on: application of methodology of a game theory and multi-criteria analysis into strategic analysis and decisions and negotiations support on competitive telecommunications market. He is employed as a Research Assistant at National Institute of Telecommunications.

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

Paper

Problems of Broadband in Rural Areas in Light of the BReATH Project Experiences

Paweł Białoń

Abstract—Some lessons learned from the EU project "Broadband e-Services and Access to the Home" (2005–2007) are presented concerning the broadband development in rural areas. In particular, the paper discusses the common problems of broadband deployment in the rural environment, various aspects of stimulating demand for broadband, the limitations of public aid and, most importantly, the problems of techno-economic analysis.

Keywords— broadband initiatives, demand, rural area, techno-economic analysis, WiMAX.

1. Introduction

This paper presents selected results of the research done within the framework of the EU project BReATH (Broadband e-Services and Access to the Home), which has run from 2005 to 2007. The project involved partners from several EU countries, among others – the National Institute of Telecommunications (NIT) from Poland. The project aimed at transferring know-how and good practices in broadband (BB) deployment to EU new member states (NMSes).

The scope of the research was rather wide. The project did not surround itself to rural areas. The research activities varied from techno-economic analyses of broadband deployment in selected areas in Europe, done on the base of mathematical simulation models to sharing experience of various broadband market players (the incumbent operator, their competitors, the local authorities, the regulator, research institutions, etc.) during numerous workshops in the partners' countries and other discussion fora like special interest groups, meetings with local authorities. Thus the outcome of the project is rather rich and the purpose of this paper is to discuss the lessons learned with respect to a particular, narrow area. Namely, we shall concentrate on the Polish specificity. Also, we shall direct our considerations to rural areas. There, the problem of broadband development is most crucial due to natural technical and economic difficulties and, consequently, the threat of digital-divide is most real.

Due to the summary character of this paper the author will often refer to more exhaustive papers and documents in order to give a more precise discussion or substantiation of the claims. It must be also stressed that some of the claims do not come from the author's own work or experience, they might be, for example, opinions expressed repeatedly by broadband professionalists during workshops and, as such, seemed valuable to the author of this paper.

The BReATH project will be shortly summarized in Section 2. Section 3 reports selected findings from the project, grouped in subtopics (like demand stimulation) techno-economic analysis made by the National Institute of Telecommunication during the project and lessons learned from it are described in Section 4. The whole paper is summarized in Section 5.

2. The Outline of the BReATH Project

The BReATH project [1] was constituted by the following 7 partners: Eindhoven University of Technology (the Netherlands), Research and Education Society in Information Technology (Greece), Gtel Consultancy Limited (UK), National Institute of Telecommunications (Poland), Institut Jozef Stefan, University of Ljubljana (Slovenia), Institute of Photonics and Electronics (Czech Republic), European Institute for Research and Strategic Studies in Telecommunications GmbH (Germany), accompanied by about 20 affiliated partners – companies, institutions, government bodies, supporting the research.

The project's goal, to transfer know-how and good practices in broadband deployment to NMSes, was realized by various research activities:

1. Benchmarking the current status of broadband development in the 7 partners' countries, including the broadband infrastructure, the internet services using broadband, the legal circumstances [2], [3]. Also, the survey of Internet access technologies was made.

2. The analysis of best practice case studies, i.e., the outstanding successful broadband initiatives; with the emphasis on the business models which had been taken [4]. The broadband development, either in Europe or outside it, seems not to have a form of a monolithic process, conforming to a single theoretical model or some regulation act. Rather, it is formed by individual initiatives of communities, cities, regions, programs of various spread, etc. Thus analyzing case studies is of utmost importance to our understanding of broadband development.

3. For some of the case studies also illustrative *techno-economic analyses* was performed, that assess their financial viability and approximate the amount of the necessary network equipment [5]. Such analyses use precise mathematical models, usually simulation

models. One of the analyses, performed by the National Institute of Telecommunications for the city of Łódź (Poland) and its surroundings, will be further reported, not only to show the results regarding the particular area but also to describe the process of performing techno-economic analysis, the problems which are then met, and the profits from having done such analysis.

4. Exchange of experiences of various actors of the broadband market. Having stated that the broadband development hardly conforms to a single theoretical model, we appreciated such experience. Workshops have been organized in the project partners' countries. They gathered scientists, representatives of the incumbent operators, other telecommunication companies, authorities (both the regulator and local authorities), non-governmental organizations (NGOs). Two of the workshops were organized in Poland [6], [7] and we shall mainly refer to the findings gathered from them. One of the workshops was much appreciated by the project partners for having revealed the important difficulties in broadband development. The interaction between various broadband players also took other forms: specially created Internet groups, cooperation with particular local authorities (e.g., Łódź, Podlasie).

5. Preparation of instructive materials for authorities willing to undertake a broadband deployment initiative, together with the *road maps* for particular partners' countries that suggests that some levels of several broadband development indicators (penetration, and others) should be attained at the specified points in the future [5], [8].

3. Problems and Solutions for Rural Areas

3.1. General Problems

Specialists name the common problems of broadband rollout in rural areas [6]. The low population density often prevents the operators from making investments there, since no viable business case is possible. Local authorities may have limited technical background, as well as too limited financial and organizational possibilities to conduct broadband initiatives.

Low density is not only a business obstacle but also a technical obstacle. Let us take an example of one the most natural technologies for rural areas, i.e., the digital subscriber line (DSL) technology, that may require only moderate changes to the existing copper telephone infrastructure in order to make it also possible to transmit digital data. Due to low population density, subscribers must be fairly distant from the exchanges. However, the more is the distance, the less band can be offered – see Table 1 (repeated after [3]). With the radio WiMAX (worldwide interoperability for microwave access) technology [9], similarly, the more distant from the base station the subscriber is, the less band the subscriber can use. In both the cases the reason is that with the growth of the distance, the signal attenuation grows and we must switch to more and more robust modulation types.

Table 1
Asynchronous DSL (ADSL) transmission reach versus bandwidth for 0.5 mm copper cable

Band (up/down) [kbit/s]		Reach [km]	
ADSL	ADSL G.Lite	ADSL	ADSL G.Lite
2048/16	64–1563/ 32–512	4.8	5–6
4096/160	–	4.0	–
6144/384	–	3.7	–
8129/640	–	2.7	–

In Poland, the relatively big land masses, together with the heritage of the monopolist position of the incumbent operator cause an observable neglect of the telecommunication access infrastructure, expressing in low broadband penetration in villages (Table 2), and even some neglect in voice services.

Table 2
Presence of Internet access in households in July 2004 [3], [10]

Space	Broadband (over 128 kbit/s)	Analog modem
Total	32%	36%
Cities	37%	35%
Villages	9%	44%

In turn, Poland has a quite well developed skeleton optical network [3]. Moreover, we should not overlook the cost barriers to broadband on the customer side. Beside the service monthly Internet access fee, we should take care about the costs of a computer set itself which, being similar to the average salary, might be a barrier to some households. Local authorities, which often have also a neglected road or water supply infrastructure, prefer improving it instead of rolling out broadband.

On the technology side, we must point at an interesting variant of DSL, used by Polish incumbent. Its reach is up to 8 km from a switchboard. However, it has a band up to 115 kbit/s, so it is disputable whether it can be called broadband (today's broadband definitions require at least 128, 256, or even 512 kbit/s). Thus, WiMAX solutions seem more prospective.

The Cornwall case – an interesting example of broadband initiative for rural areas was presented during a BReATH workshop [7, presentation by J. Cowans]. This example will be present throughout the rest of this section. Now we call this example in order to show that the neglect of some rural areas by Internet providers is neither the specificity of Poland, nor of NMSes.

Cornwall is a peripheral, rural county in UK, geographically isolated. It has half a million inhabitants.

It used to be a mining area, producing china clay, tin and copper. Through the 20th century mining has declined.

In 1999 its market position was limited, the gross domestic product (GDP) was below 60% UK average and below 70% of EU average; the emigration became a problem. Cornwall was then not only poorly interconnected but also had no prospect of being on any broadband operator's map. Moreover, the computer skills of the inhabitants were poor.

3.2. Problems with EU and Other Public Aid

Public aid, including EU structutral funds, seems a natural remedy for areas where private investments in broadband seem not profitable. However, several drawbacks of such aid in its current state have been pointed during the workshops [6], [7]:

- Mr Sadowicz stated that the structural funds in Poland, which were intended to lower the digital divide, may in practice increase it, since they prefer large projects [6]. Mr Sadowicz with his organization Cities on Internet investigated the use of funds from *Integrated Regional Development Operational Programme, Measure 1.5* (directed to building the information society). In the period of 2004–2006 over 202 broadband projects were granted funds by the programme. In general, poorer regions got less or smaller grants (with the difference approaching PLN 42 million = EUR 10.5 million, which already exceed the total budgets of some regions). Moreover, only 12 of the projects could be considered "authentic" in that they do not limit themselves to particular beneficiaries, like public administration, schools, etc. These "authentic" projects were run only in half of the regions.

- The usual complains about procedures complexity and inability of some local governments to apply for structural funds were rosen.

- Using the existing regions in allocating funds may be irrelevant. One from the audience, prof. T. Grzeszczyk, gave an example of the Mazovia region. As containing the capital of Poland, it has a relatively big GDP, on average, thus its position while applying for public aid is weak. But it is enough to go a few tens kilometers away from Warsaw to observe one of the deepest poverty levels in Poland.

- Some hope for rural regions may be attributed to the European Commission's concept of universal service [7, presentation by P. Kenney]. With this concept, where the market fails to assure broadband connections, an intervention could take place (whose framework cost would amount to 50 euro cents per EU citizen). However, at least until 2009, this concept is not in power and countries are not very enthusiastic about introducing it.

The remedies must be found to all the above problems; some hope is connected NGOs supporting the local authorities in acquiring public funds (like Cities in Internet and Information Society Building in Rural Areas – e-Vita, mentioned later).

3.3. Stimulating Demand as a Key to Success

One of the most important findings of the BReATH project was the importance of the economic demand stimulation. Instead of funding building the network infrastructure by local authorities (or consortia including them) we can have it done by the network operators/providers present on the market. However, to make their business case viable, we could increase the demand by donating the broadband subscribers with the use of public funds. Directing public funds to the subscribers instead of the operators seems to less disrupt the market mechanisms. More importantly, such a concept is firmly supported with practical cases.

The Cornwall case success lies – according to their authors – in a suitable demand stimulation. The ACTNOW project, that brought broadband to Cornwall, was started in 1999 and was a sort of public-private partnership, using EU funds (the main stakeholders were, except of the authorities, the incumbent, British Telecom (BT), the Regional Development Agency, the Business Link company). ADSL was chosen as the most viable technology at that time.

The main idea of the project was to stimulate broadband demand by subsidizing small and medium enterprise (SME) connections. The SMEs that subscribed to broadband were for two years donated 50% of their subscription costs. Extensive campaigns were conducted to convince the SMEs that broadband is what can leverage their businesses. A part of these campaigns was monitoring local GDP and some other business activity merits (also, there were voices for directing public funds to other goals: building roads, perhaps an airport or a university and the local community was persuaded that broadband would give better economic effects). The businesses were given also an extensive support in using services.

The project amounted to £12.5 million has been run in several stages. The targets of covering 50% of businesses and modernizing exchanging offices have been exceeded with a business coverage of 99%. Above 5000 jobs have been created and safeguarded and the overall take up in Cornwall has reached 30%, which exceeds the national average of 25%. The 80% of the surveyed businesses now consider broadband critical for their activities. Most Cornwall business employ 2–5 people. There is higher employment, inward investment, and even some immigration.

Some similar case is the case of Polish community of Stoszowice [6, presentation by Patryk Wild]. Stoszowice is located in the Lower Silesia region. With 5700 inhabitants and 28% unemployment in 2004 it has one of the lowest tax incomes in the area. Sewage and water supply systems were lacking but the mountainous and rural landscapes, and unpolluted environment have potential of attract tourists. The TP SA ignored Stoszowice in their broadband development strategy, Stoszowice did not possess any Internet connection.

The initiative of Stoszowice was co-funded by the Ministry of Science and Information and by the partners of the e-Vita programme, Cisco Systems, the Rural Development Foundation and the Polish-American Freedom Foundation. The e-Vita and Cisco brought also a considerable technological and organizational know-how inside to the initiative.

The demand was stimulated by making the Internet access free of charge for 2 years. Patryk Wild, the president of Stoszowice, who is a young and vivid person, was able to convince the citizens the profits from broadband to the extent that they preferred investing broadband over investing in roads. The network has been promoted via community newspaper and community meetings. A communal educational portal has been built and is still being developed. It is directed to parents of kids from communal schools and kindergartens as well as to teachers and pupils.

The access infrastructure was built by various subjects, including the incumbent, using various technologies (WiFi IEEE 802.11b, ADSL). In 2006, more than 75% of the communal households were within the range of the WiFi (wireless fidelity) network. Over 110 families (out of the community total of about 1600) used the Internet. There were then about 100 new orders for connections to the network. An extension of the network was planned so that it would reach about 95% of all households in the community.

Both local and national press gave very positive comments about the initiative.

In Poland, the *law on development of information technologies for subjects realizing public tasks* from 17 February 2005, had an extreme potential of stimulating demand on broadband throughout the whole country. It had a revolutionary character in enforcing introducing e-government services to numerous government institutions. The law was also to impose the creation of necessary, uniform data formats for such services. Unfortunately, before the law came into force, it became clear that its broad plans would remain only "on paper", since the government was not able to fulfill the law and concrete actions were not taken.

3.4. The Role of Social Leaders

In demand growth it is important that some future broadband consumers learn to use it from the current users (the later are called social leaders). In many cases we can observe the activity of social leaders:

- The conductors of the Cornwall initiative claim to have observed an increased broadband take-up in households after the businesses had signed for broadband.

- In one of the biggest projects in Poland, Broadband Communication Network of Kuyavia and Pomerania Region [11], the role of social leaders is attributed to universities, libraries, hospitals, etc., who were the first customers of the network being built.

- The e-Vita treat their successful initiatives as social leaders for future initiatives, that's why they describe/disseminate their successful cases.

4. Techno-Economic Analysis

We shall describe the techno-economic model and analysis performed for the city Łódź and its rural surrounding. However, even more important than particular results for Łódź will be the general remarks regarding problems of performing techno-economic analysis and the potential role of research institutions.

4.1. The Importance of Techno-Economic Analysis

A techno-economic analysis (TEA) aims at dimensioning the network, meaning assessing the necessary number of particular equipment items and at assessing the profitability of the undertaking under various economic and market assumptions. The profitability assessment uses various financial indicators like internal rate of return, net present value, cash balance, payback period. Thus TEA seems a natural component of any broadband initiative undertaken by local leaders. Then, however, a problem arises to whom to entrust the TEA.

The City Hall of Łódź observed that it was difficult to find a company willing to perform a TEA that would be really independent on any operator – a potential contractor of their initiative. In general, not many can perform a TEA. The necessary TEA know-how is a pretty well protected knowledge. Publications, like [12]– [14] or [15] are rare and many solutions are based on particular subjects' knowledge. An example is the Titan tool used in the analysis of the broadband development in Rome [14], or the STEM series of tools. The team of NIT agreed to build a techno-economic model for Łódź and to perform an independent TEA, based on the broad expertise of the institute in telecommunication networks gained during the several decades of the research. The TEA was done within the framework of BReATH project.

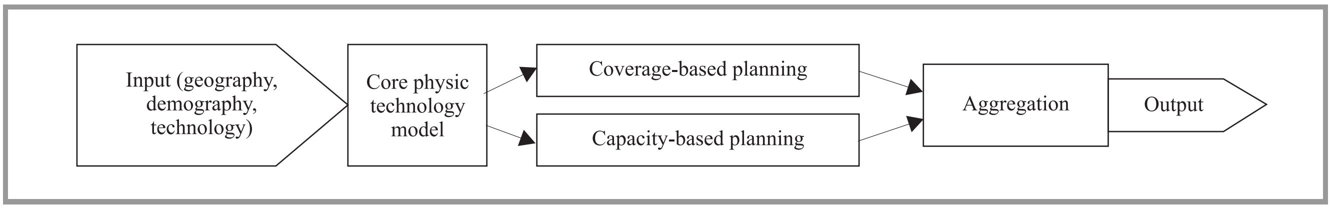

Input (geography, demography, technology)
Core physic technology model
Coverage-based planning
Capacity-based planning
Aggregation
Output

Fig. 1. Technological model.

4.2. The Techno-Economic Analysis for the City of Łódź and Surroundings

Our TEA and the model are described thoroughly in [5], [16] and [17].

Łódź is the second largest city in Poland. It has well distinguished urban, suburban and – recently incorporated – rural regions. The main network infrastructure was to remain the property of the town. It can be thus the subject of our analysis. However, serving customers was to be done by private companies. These companies should be responsible for added-value, such as security and the software side of the Internet services. Thus, there arose one more question: how the revenue should be divided between the infrastructure owner and the service providers.

The city has been split into 3 areas urban (a), suburban (b) and rural (c), an further – into 20 smaller ones, for which we obtained data about the number of inhabitants and several types of businesses/institutions from the Statistical Office.

The WiMAX technology was chosen for areas a, b and c in our analysis, due to its simplicity in terms of the civil work needed and the need to gain independence of the incumbent operator.

Our TEA was based on the simulation model we build (with elements of inverse simulation, i.e., solving equations and elements of the constrained analysis). The model parts (demand model, technological model, economic model) will be now described.

The **demand model** calculates the demand, i.e., the expected number of users categorized by:

- user types (individual, companies, schools, municipality offices),
- bandwidths,
- areas.

The demand model can be decomposed into 5 main processes and uses input data coming from diverse sources, varying from the Statistical Office to heuristic observations, own expertise of the NIT and results of other researches, often collected through the Internet.

The **technological model** (Fig. 1) calculates the necessary amount of equipment (mainly base stations and their sectors, consumer premise equipment – CPE). It follows the work [15] and adopts the path loss model from [13].

The part depicted as "core physics and technology model" describes the physical phenomena taking place in the network. The licenced frequency of 3.5 GHz was chosen, with the bandwidth of 56 MHz, which be divided into channels of either 3.5 or 7 MHz. The model takes account of number of sectors, radiation angles, transmission powers and gains, the sensitivities of CPE. *Link budgeting* is used in the model, i.e., the loss of a logarithm of signal power is calculated. This loss is a sum of several components connected with the distance from the base station, terrain attenuation, gain from emitter angle, etc. For example, the terrain attenuation is defined as

$$A - B \cdot h_b + C/h_b\,,$$

where h_b is the height of base station installation and A, B, C – the terrain coefficients, each of them defined differently for three different terrains: a, b, c. Signal losses connected moving people, tree leaves, cars passing are computed from a probabilistic model using lognormal distributions.

The most important decision made during the calculation is the choice of modulation type, since it changes the trade-off between the necessary sensitivity of the receiver and a channel width (Table 3) and – in consequence – influences the trade-of between the coverage and capacity of our network. We compute the minimal possible number of base stations (together with the modulation type) such that both the constraints on coverage and on capacity are satisfied.

Table 3
Sensitivities of the receivers [dBm]

Channel width [MHz]	Modulation type					
	QPSK 1/2	QPSK 3/4	16-QAM 1/2	16-QAM 3/4	64-QAM 2/3	64-QAM 3/4
1.75	–90	–87	–83	–81	–77	–75
3.5	–87	–85	–80	–78	–74	–72
7	–84	–82	–77	–75	–71	–69
14	–81	–79	–74	–72	–68	–66

QPSK – quadrature phase shift keying,
QAM – quadrature amplitude modulation.

The **economic model** simulates the cash flows during the project duration. It is a dynamic model. The time horizon has been set to 7 years, as this seems the longest time the telecommunication companies want to wait for investment payback; otherwise they do not make the investment. Also, economic, market and demographic prognoses exceeding a 7 year horizon are certainly not much reliable.

There are 4 outputs of the model: cash balance and net present value – NPV (both scalar functions of time) and the payback period and internal rate of return – IRR (both scalars). We assume that the project starts at time $t = 0$ and that time is measured in months (i.e., $t = 1$ means the time 1 month after the start of the project).

The definitions of economic quantities taken in this model are the following:

- Capital expenditure (CapEx): the sum of money spent on investments (assumed to be made once, at the beginning).
- Operational expenditure (OpEx(t)): the sum of money spent in the time interval $(t, t+1]$. It includes only money that is paid periodically to enable running a project (e.g., maintenance costs).
- Cash balance (t): the sum of revenues from the period $(0, t+1]$ minus sum of expenditures from period $(0, t+1]$. All the flows are taken in their nominal values, in which they are booked.
- Net present value, NPV(t): a measure similar to cash balance (t), taking into account, however, that money depreciates in time: equal nominal values correspond to different real (discounted) values. More precisely, with a monthly money depreciation rate r, NPV is defined as follows:

$$\mathrm{NPV}_r(t) = \sum_{u=1}^{t} \left(\frac{\text{cash balance}(u) - \text{cash balance}(u-1)}{(1+r)^{u-1}} \right).$$

- Payback period: the minimal time t such that NPV$(t) = 0$. Usually, due to capital investments made, NPV is initially below zero and it grows during the project run:

$$\text{payback period} = \min(t | \mathrm{NPV}(t) \geq 0).$$

- Internal rate of return: r for which $\mathrm{NPV}_r(t) = 0$. Comment: we treat our business as investing money in a bank and getting them back. The "surrogate bank rate" is our IRR.

The economic model takes into account customers' fees, subsidies, the cost of installing and maintenance of equipment, taxes, depreciation of the equipment. The customers are assumed to pay a monthly fee and no initial (installation) fee. A constant part of revenues goes to the service operators (this part is a model parameter, called end operators' profit margin) and the rest goes to the city.

Table 4
Influence of the subsidies to IRR and payback period

Subsidy [million PLN]	0	7	15	30
Payback period [month]	46	39	31	20
IRR (monthly) [%]	2.25	2.76	3.50	5.92

Analysis – many runs of the model has been done [5], [17], for various terrains (a, b, c) and various assumptions regarding the values of market share by our network, amount of subsidy, end-operators' margin profit and others. Here we shall only call the results for the experiment most relevant to our subject. This experiment conceived the rural (c) part of Łódź and showed the influences of the subsidy to the economic indicators. The results are given in Table 4 and Fig. 2. In the rural area of Łódź there are about 10,000 households, 4600 businesses (most of them being small businesses), 1000 offices, schools and other public institutions. The rural area of Łódź is 151 km^2.

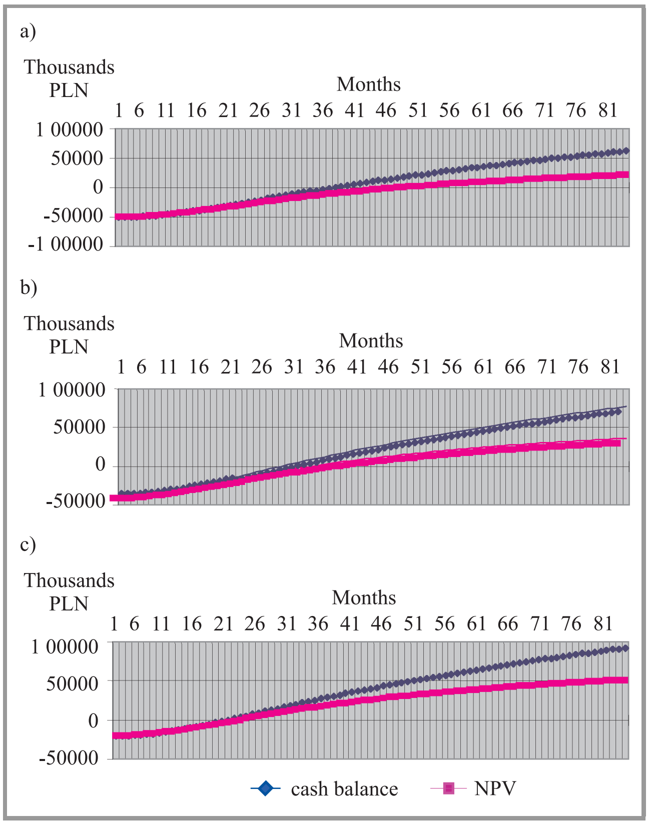


Fig. 2. Influence of the subsidies to cash balance and NPV. Subsidy: (a) 0 PLN; (b) 15 million PLN; (c) 30 million PLN.

The results seem somewhat astonishing, since even under the absence of subsidies the investment pays back after 3 years. The common experience says for rural areas the investments do not pay back. This was explained with the fact that the rural surroundings of Łódź, not typically, have a high ratio of institutional subscribers among all the subscribers (and institutions are known good clients due to high prices for their high bandwidth and high quality Internet connections). Subsidies, however, have a potential of essentially shortening the payback period (increase the IRR).

4.3. Problems of Techno-Economic Analysis

Our experience with TEA for broadband initiatives showed a need of undertaking systematic research in this area by research institutions of various specialties (telecommunication, economy, demography, etc.). The know-how in TEA is not public knowledge, and as broadband initiatives need independent techno-economic analysis, they could refer to research institutions. They, however, require a great amount of detailed information and firm experience that could only be gained in systematic research, perhaps in cooperation.

Let us name the problems we met during our analysis to see how specialized knowledge is necessary. Collecting necessary data was extremely difficult. Some coefficients (A, B, C) expressing the terrain attenuation (by plants, cars, people) are obviously very hard to obtain. Foreseeing a demand generated by particular was often more art than knowledge. It turned out that the Polish Statistical Office does not make statistics with granularity less than a single town: obtaining data for our 20 pieces of Łódź required a dedicated research by them. The lack of the supporting literature has already been mentioned.

5. Conclusions

The BReATH project brought many interesting findings about developing broadband in communities; the reach material from this project can be further referred. The advantages of the technique of stimulating demand market demand for broadband connections were stressed. The role of social leaders is essential. The main domain for the activity of research institutions seem to be techno-economic analysis, which, however, requires a period of creating the necessary knowledge, in a cooperative research.

Acknowledgments

The work on the techno-economic model and analysis has been performed jointly by Paweł Białoń, Agnieszka Gosk, Janusz Granat and Marcin Salwa from the NIT, within the framework of the BReATH project. The research was co-funded by two Polish Ministry of Science grants. We would like to thank Mr Jacek Zieliński from the City Hall of Łódź for his cooperation and Mr Stanisław Kaniewicz, Mr Sławomir Kończak and Mr Jacek Sitkiewicz from the Statistical Office in Łódź for acquiring necessary data. Special thanks go to Mr Goff Hill (GTel Consulting, UK) for his valuable comments on the model made during its construction. Some findings reported here have been elaborated by tens of people participating, more or less closely, in the BReATH project.

References

[1] “The BReATH project” [Online]. Available: http://ist-breath.net

[2] “BReATH Deliverable D2.1: Preliminary survey report”, 2005-09-12 [Online]. Available: http://ist-breath.net

[3] “BReATH Deliverable D2: Surveying study report – general sections”, 2006-01-16 [Online]. Available: http://ist-breath.net

[4] “BReATH Deliverable D3.2: First set of case studies, detailed data, analysis and results”, 2006-09-12 [Online]. Available: http://ist-breath.net

[5] “BReATH Deliverable D3.5: Final survey of best broadband access deployment practices and solutions”, Part 3: “Techno-economic case studies”, 2007-02-28 [Online]. Available: http://ist-breath.net

[6] “BReATH Deliverable D1.1: Material gathered from open event E1, workshop in Poland”, 2006-06-02 [Online]. Available: http://ist-breath.net/events/eventE1.html

[7] “BReATH Deliverable D1.6, Part 1: Material gathered from open event E6, follow-up workshop in Poland”, 2007-01-13 [Online]. Available: http://ist-breath.net/events/eventE6.html

[8] “BReATH Deliverable D5.1: Road map report”, 2007-02-19 [Online]. Available: http://ist-breath.net

[9] WiMAX Forum: “WiMAX deployment considerations for fixed wireless access in the 2.5 GHz and 3.5 GHz licenced bands”, 2005 [Online]. Available: http://www.wimaxforum.org

[10] „Wykorzystanie technologii informacyjno-telekomunikacyjnych w przedsiębiorstwach i gospodarstwach domowych w 2004 r.”, Warsaw: Główny Urząd Statystyczny (in Polish) [Online]. Available: http://www.stat.gov.pl/dane_spol-gosp/spoleczenstwo_informacyjne/2004/index.htm

[11] “The Broadband Communication Network of Kuyavia and Pomerania Region” [Online]. Available: http://www.kpsi.pl

[12] A. Gumaste, I. Chlamtac, and C. A. Szabó, *Broadband Services. Business Models and Technologies for Community Networks*. Chichester: Wiley, 2005.

[13] V. Erceg *et al.*, “Model of wireless channels in suburban enviroments”, *IEEE J. Selec. Areas Commun.*, vol. 17, no. 7, pp. 1205–1211, 1999.

[14] C. Kolias, A. Makris, N. Nikopoulos, P. Panagiotopoulos, I. Zacharopoulos, G. Yovanof, and I. Tomkos, “Comparative study and techno-economic analysis of residential broadband access technologies: GPON and FWA”, *J. Commun. Netw.*, no. 4, pp. 211–217, 2005.

[15] T. Smura, “Techno-economic analysis of IEEE 802.16a-based fixed wireless access networks”. Ph.D. thesis, Helsinki University of Technology, 2004.

[16] P. Białoń, “Techno-economic model and analysis of broadband deployment in the city of Łódź with the WiMAX technology”, *J. Instit. Telecommun. Profess.* (former *J. Commun. Netw.*), vol. 1, no. 7–9, pp. 41–46, 2007.

[17] J. Granat, P. Białoń, A. Gosk, and M. Salwa, “Zarządzanie usługami informatycznymi oraz infrastrukturą informatyczną”. Część B: “Modele wspierające rozwój regionalnych sieci dostępowych”, Report 06.30.002.6. Warsaw: National Institute of Telecommunications, 2006 (in Polish).

Paweł Białoń – for biography, see this issue, p. 39.