JOURNAL OF TELECOMMUNICATIONS
AND INFORMATION TECHNOLOGY

3/2010

Editorial Board

Editorial Advisory Board

ISSN 1509-4553 on-line: ISSN 1899-8852

Warsaw 2010

Circulation: 300 copies

Sowa - Druk na życzenie, www.sowadruk.pl, tel. 022 431-81-40

JOURNAL OF TELECOMMUNICATIONS
AND INFORMATION TECHNOLOGY

Preface

This issue contains ten papers, four of which deal with multi-objective optimization problems in contemporary telecommunications, the remaining six with diverse problems of knowledge engineering or optimization related to telecommunications or information society.

Teresa Gomes and José Craveirinha in the paper *An Algorithm for Enumerating SRLG Diverse Path Pairs* start with the premise that telecommunication networks are intrinsically multilayered, a single failure at a lower level usually corresponds to a multi-failure scenario at an upper layer. In this context, the concept of shared risk link group (SRLG) diverse path set may be defined as a set of paths, between an origin and a destination, such that no pair of paths can be simultaneously affected by any given failure (or risk) in a single failure scenario. Firstly we present the formulation of the SRLG diverse path pair calculation problem in a directed network. An algorithm for enumerating SRLG diverse paths, by non decreasing cost of their total (additive) cost will be presented, which is based on an algorithm proposed for generating minimal cost node disjoint path pairs. The SRLG diverse path pairs may be node or arc disjoint, with or without length constraints. Computational results are presented to show the efficiency of the proposed algorithm for obtaining node or arc disjoint SRLG diverse path pairs in undirected networks.

Carlos Simões, Teresa Gomes, José Craveirinha, and João Clímaco in their paper *Performance Analysis of a Bi-Objective Model for Routing and Wavelength Assignment in WDM Networks* recall that establishing end-to-end connections on wavelength division multiplexing (WDM) networks requires setting up lightpaths, defining the sequence of optical fibres and the wavelength in each fibre (the routing and wavelength assignment problem) for traffic flow. This paper reviews a bicriteria model for obtaining a topological path (unidirectional or symmetric bidirectional) for each lightpath request in a WDM network, developed by the authors, and presents a performance analysis of the model by considering important network performance measures. A performance analysis of the two bicriteria model is presented, comparing the performance metrics obtained with the monocriterion models using the same objective functions, in five different standard reference networks.

The same authors in the paper *Performance Analysis of a Bi-Objective Model for Routing with Protection in WDM Networks* add the issue of a fault management scheme in WDM's in order to avoid the huge loss of data that can result from a single link failure. Dedicated path protection, which establishes two end-to-end disjoint routes between the source-destination node pair, is an effective scheme to preserve customers' connections. This paper reviews a bicriteria model for dedicated path protection, that obtains a topological path pair of node-disjoint routes for each lightpath request in a WDM network, developed by the authors. A performance analysis of the bicriteria model is then presented, comparing the performance metrics in a similar setting as in the previous paper.

Michał Marks in the paper *A Survey of Multi-Objective Deployment in Wireless Sensor Networks* addresses the problem in designing wireless sensor networks (WSNs) that relates to finding a tradeoff between the desired requirements for the lifetime, coverage or cost of such a network while coping with the computation, energy and communication constraints. The paper examines the optimal placement of nodes for a WSN in a multi-objective formulation. It is impossible to consider the deployment of the nodes separately from WSNs applications. The properties of WSNs applications that determine the placement problem are highlighted. Diverse objectives that should be considered are defined and enumerated. The paper provides an overview and concentrates on multi-objective strategies, their assumptions, optimization problem formulations and results.

Cezary Chudzian and Jarosław Sobieszek in the paper *Personal Ontologies for Knowledge Acquisition and Sharing in Collaborative PrOnto Framework* summarize preliminary experiences with a prototype framework for collaborative knowledge acquisition and sharing, called PrOnto. At the moment the artifacts that are organized and shared are typical sources of scientific knowledge, namely journal papers and web pages. In PrOnto two interrelated explicit levels of knowledge representation are introduced: keywords and ontological concepts. Each user of the framework maintains his own ontological profile, consisting of concepts and each concept is, in turn, by subjective user's decision, related to a set of weighted keywords that define its meaning. Furthermore, dedicated indexing engine is responsible for objectively establishing correspondence between documents and keywords, or in other words, the measure of representativeness of the keyword to document's content. Developing an appropriate knowledge model is a preliminary step to share it efficiently. The higher level representation facilitates exploration of other people's areas of interest. PrOnto gives an opportunity to browse knowledge artifacts from the conceptual point of view of any user registered in the system. The paper presents the ideas behind the PrOnto framework, gives an outline of its components and finalizes with a number of conclusions and proposals for future enhancements.

Ewa Niewiadomska-Szynkiewicz and Michał Marks in the paper *A Software Platform for Global Optimization* address issues associated with the global optimization algorithms, which are methods of finding optimal solutions for complex (non-convex, discrete, etc.) problems. The paper focuses on an integrated software environment – global optimization object-oriented library (GOOL), which provides the graphical user interface together with the library of solvers for convex and nonconvex, unconstrained and constrained, although mostly non-discrete problems. The design, performance and possible applications of the GOOL system are described. A practical example – price management problem – is provided to illustrate the effectiveness and range of applications of the software tool.

Agnieszka Gosk in the paper *Query Optimization in Teradata Warehouse* presents a definition of the active data warehousing (ADW) paradigm. One sort of data warehouses which is consistent with this paradigm is teradata warehouse. Therefore, the basic elements of the teradata architecture are described, such as processors parsing engine (PE) and access module processor (AMP). Emphasis was put on the analysis of query optimization methods. The impact of a primary index on the time of query execution is discussed. Furthermore, the paper shows different methods of optimization of data selection, data joins and data aggregation. All these methods can help to minimize the time for data processing. The paper presents experiments which show the usage of different methods of query optimization. Conclusions about different index usage are included.

Paweł Białoń in the paper *Solving Support Vector Machine with Many Examples* presents and compares various methods of dealing with linear support vector machine (SVM) problems with a large number of examples. The author believes that some interesting conclusions from this critical analysis apply to many new optimization problems and indicate in which direction the science of optimization will branch in the future. This direction is driven by the automatic collection of large data to be analyzed, and is most visible in telecommunications. A stream SVM approach is proposed, in which the data substantially exceeds the available fast random access memory (RAM) due to a large number of examples. Formally, the use of RAM is constant in the number of examples (though usually it depends on the dimensionality of the examples space). It builds an inexact polynomial model of the problem. Another approach is exact. It also uses a constant amount of RAM but also auxiliary disk files, that can be long but are smartly accessed. This approach bases on the cutting plane method, similarly as Joachims' method (which, however, relies on early finishing the optimization).

Wojciech Szynkiewicz in the paper *Planning System for Multi-Agent Based Reconfigurable Fixtures* describes a concept of the planning system for self adaptable, reconfigurable fixtures composed of mobile locators (robotic agents) that can freely move on a bench and reposition below the supported part, without removing the part from the fixture. The main role of the planner is to generate the admissible plan of relocation of the mobile agents. A constrained nonlinear optimization problem is formulated to find the optimal locations for supporting heads.

Mahmoud Youssuf and Mohamed Z. Abdelmageed in the paper *Performance Analysis of Hybrid Phase Shift Keying over Generalized Nakagami Fading Channels* consider hybrid phase shift keying (HPSK) modulation that reduces the peak to average power ratio of the transmitted signal, reduces the zero crossings and the $0°$-degree phase transmissions, but also enhances the reduction of the bit error rate (BER) measure of the signal performance. The properties of HPSK are analyzed, and an expression for the conditional probability of HPSK modulation over additive white Gaussian noise (AWGN) is derived. This BER measure of HPSK is shown to outperform quadrature phase shift keying (QPSK) modulation. HPSK performance through a generalized Nakagami fading channel is also considered.

We wish our Readers an interesting reading time.

Andrzej P. Wierzbicki
Guest Editor

Paper

An Algorithm for Enumerating SRLG Diverse Path Pairs

Teresa Gomes and José Craveirinha

Abstract—Telecommunication networks are intrinsically multi-layered, a single failure at a lower level usually corresponds to a multi-failure scenario at an upper layer. In this context, the concept of shared risk link group (SRLG) allows an upper layer to select, for a given active path (AP), a backup path (BP), which avoids every SRLG that may involve the selected AP, in the event of a failure. That is a SRLG diverse path set maybe defined as a set of paths, between an origin and a destination, such that no pair of paths can be simultaneously affected by any given failure (or risk) in a single failure scenario. Firstly we present the formulation of the SRLG diverse path pair calculation problem in a directed network. An algorithm for enumerating SRLG diverse paths, by non decreasing cost of their total (additive) cost will be presented, which is based on an algorithm proposed for generating minimal cost node disjoint path pairs. The SRLG diverse path pairs may be node or arc disjoint, with or without length constraints. Computational results will be presented to show the efficiency of the proposed algorithm for obtaining node or arc disjoint SRLG diverse path pairs in undirected networks.

***Keywords*—routing, SRLG disjoint shortest paths, telecommunication networks.**

1. Introduction

Bandwidth usage optimization is one of the main issues when protection schemes are used in telecommunication networks. In global path protection, the path that carries the associated traffic flow under normal operating conditions is called the active path (AP), and the path that carries that traffic when some failure affects the AP is called the backup path (BP).

Many network providers consider sufficient to implement protection schemes which ensure their network (or certain connections in their network) is 100% reliable in single failure scenarios. Because telecommunication networks are intrinsically multi-layered, a single failure at a lower level usually corresponds to a multi-failure scenario at an upper layer.

A failure risk may represent a fibre cut, a card failure at a node, a software failure, or any combination of these factors [1], which may affect one or more links at a given network layer. In this context, the concept of shared risk link group (SRLG) is very important in teletraffic engineering since allows an upper layer to select, for a given AP, a BP, which avoids every SRLG that may involve the selected AP, in the event of a failure. Note that this may not be feasible for all possible APs. That is a SRLG diverse path set maybe defined as a set o paths, between an origin and a destination, such that no pair of paths can be simultaneously affected by any given failure (or risk) in a single failure scenario. Therefore, to ensure global path protection against a single failure affecting a single SRLG, a SRLG diverse path pair must be calculated.

The problem of finding a SRLG diverse path pair has been shown to be NP-complete [1]. The minimum-cost diverse routing problem, in which the objective is finding two paths, SRLG diverse, with minimal total arc cost (also designated as the min-sum problem), is also NP-complete [1]. Hu [1] proposed an integer linear programming (ILP) formulation for the min-sum problem, and provide numerical results showing that the ILP formulation quite effective in networks with a few hundreds of nodes.

The necessity of calculating SRLG diverse path pairs arises in optical and multiprotocol label switching (MPLS) networks, where certain connections require two paths, the AP and the BP, in order to satisfy service level agreements (SLA) regarding reliability. The possibility of enumerating, by non-decreasing cost, SRLG diverse path pairs, may allow more elaborate, and possibly more efficient, forms of SRLG diverse routing. Furthermore the ordered enumeration of diverse SRLG paths will make it possible a multi-objective routing approach, with survivability requirements.

Rostami *et al*. [2] proposed an algorithm, named CoSE (conflicting SRLG exclusion), which is an extension to SRLG-disjoint routing of a link-disjoint routing algorithm called CoLE (conflicting link exclusion), proposed in [3], which can quickly find an optimal solution path pair. The CoSE algorithm iteratively separates the network SRLGs into two sets and then computes the working and backup paths. Furthermore, in [2] the authors also propose a way of calculating two maximally SRLG diverse paths in a network where no two completely-disjoint paths exist. The CoSE algorithm can be used for solving the min-min problem, by selecting the appropriate solution from the set of generated solutions (although the optimality of the solution is not guaranteed).

Todimala and Ramamurthy [4] proposed an iterative heuristic, based on a modification of Suurballe's algorithm [5], [6], for diverse routing under SRLG constraints that computes the least cost SRLG diverse paths pair. In [7] the same authors propose a heuristic for solving the problem of computing optimal SRLG/link diverse paths under shared protection (considering the definition of an optimal SRLG diverse path pair under shared protection as asymmetrically-weighted [8]).

In [9], [10] the authors consider the problem of path protection in wavelength-routed networks with SRLG and propose a heuristic method, which they compare with the trap avoidance (TA) algorithm [11]. They conclude the new algorithm, minimum total weight (MTW) algorithm, outperforms TA algorithm within the first few iterations. If more iterations are considered there is no clear advantage of one algorithm, over the other.

This work presents an exact algorithm for enumerating SRLG diverse path pairs in a multi-layered network by decreasing order of the total cost. Firstly we present the formulation of the SRLG diverse path pair calculation problem in a directed (implicitly multi-layered) network. Secondly we formalize the proposed algorithm, that is based on Algorithm 1 proposed in [12] for generating minimal cost node disjoint path pairs. The SRLG diverse path pairs may be node or arc disjoint, and with or without length constraints, as will be explained. Finally, computational results will be presented to show the efficiency of the proposed algorithm for obtaining node disjoint SRLG diverse path pairs.

The paper is organized as follows. In Section 2 the notation and the problem formulation are given. An algorithm for enumerating SRLG diverse paths, by non decreasing cost of their total (additive) cost is presented in Section 3. The application to path pairs node or arc disjoint and with length constraints, is briefly explained in Subsections 3.3 and 3.4. In Section 4 results which illustrate the algorithm efficiency in obtaining SRLG node disjoint path pairs, are presented. Finally, some conclusions are presented in Section 5.

2. Notation and Problem Definition

The algorithm in Section 3 is based on Algorithm 1 in [12]. Therefore, we will use a notation similar to the one in [12]. Let $G = (N,A)$ be a directed network with node set $N = \{v_1, v_2, \dots, v_n\}$ and arc set $A = \{a_1, a_2, \dots, a_m\}$ (where n and m designate the number of nodes and arcs in G, respectively). Let a non-negative cost function (or metric) in the arcs, be defined:

$$c_{v_a v_b} \geq 0, \quad (v_a, v_b) \in A\,, \tag{1}$$

where $c_{v_a v_b}$ represents the cost of using arc (v_a, v_b).

The cost $c(p)$ of a path p in G with respect to metric c is:

$$c(p) = \sum_{(v_a,v_b)\in p} c_{v_a v_b}\,. \tag{2}$$

Definition 1: A path p is said to be simple (or loopless) if all its nodes are different.

We will use the word path to refer to simple paths, and we will only use the expression "simple path" when required, namely in the algorithm.

Let path $p = \langle v_1, a_1, v_2, \dots, v_{i-1}, a_{i-1}, v_i\rangle$, be given as an alternate sequence of nodes and arcs of G, such that the tail of a_k is v_k and the head of a_k is v_{k+1}, for $k = 1, 2, \dots, i-1$ (all the v_i in p are different). Let the set of nodes in p be $V^*(p)$ and the set of arcs in p be $A^*(p)$. Two paths $p = \langle v_1, a_1, v_2, \dots, v_{i-1}, a_{i-1}, v_i\rangle$ and q are arc-disjoint if $A^*(p) \cap A^*(q) = \emptyset$. Two paths p and q are disjoint if $V^*(p) \cap V^*(q) = \emptyset$, and are internally disjoint [13] if $\{v_2, \dots, v_{i-1}\} \cap V^*(q) = \emptyset$. We will say that two paths are node disjoint if they are internally disjoint.

Let R be a set representing the risks (failures) in the functional network. Each risk may correspond to a fibre cut, a card failure at a node, a software failure, or any combination of these factors. Let A_r represent the subset of network arcs (or links) in the network logical representation (corresponding to a capacitated graph) that can be affected by risk $r \in R$. Thence A_r is a SRLG (associated with r).

Let

$$r_p = \{r \in R:\ \text{path } p \text{ contains elements of } A_r\}\,. \tag{3}$$

The SRLG problem can be defined as follows [1].

Definition 2: Find two paths p and q, between a pair of nodes, such that $r_p \cap r_q = \emptyset$. We also say that p and q are two SRLG diverse paths (with respect to R).

The first addressed problem is to enumerate node disjoint simple path pairs (p_i, q_i) $(i = 1, 2, \dots)$, in G, from a source s to a destination node t $(s \neq t)$, which are SRLG diverse, by non-decreasing total cost of the pair, defined by

$$c[(p_i, q_i)] = c(p_i) + c(q_i), \quad i = 1, 2, \dots\,, \tag{4}$$

where p_i and q_i have the same source and sink node.

Let R_a be the set of risks that can affect arc $a \in A$:

$$R_a = \{r : a \in A_r\}, \quad \forall a \in A\,, \tag{5}$$

R_a can be obtained from A_r $(r = 1, \dots, |R|)$ and

$$r_p = \cup_{a\in p} R_a\,, \tag{6}$$

which is much more adequate for generating SRLG diverse paths in the proposed algorithm.

If a path pair (p, q) is SRLG diverse then it is arc disjoint (regardless of whether the the network is directed or not).

Definition 3: Two arcs, $a_i, a_j \in A$ are SRLG diverse if $R_{a_i} \cap R_{a_j} = \emptyset$.

Definition 4: An arc $a \in A$ is SRLG diverse with a path p if $R_a \cap r_p = \emptyset$.

The algorithm proposed in Section 3 is based on Algorithm 1 in [12], which uses the MPS algorithm [14] in its loopless version [15]. The algorithm MPS is a *deviation* algorithm. Each time a path p is chosen from a set of candidate paths, X, new paths may be added to X. In the context of the algorithm the node v_k of path p, from which a new candidate path is generated, is the *deviation node* of that new path (which coincides with p up to v_k). In a path the link the tail of which is the deviation node, is called the *deviation arc* of that path [14]. By definition s is the deviation node of p_1 (the shortest path from s to t). The *concatenation* of path p, from v_i to v_j, with path q, from v_j to v_l, is the path $p \diamond q$, from v_i to v_l, which coincides with p from v_i to v_j and with q from v_j to v_l.

Let $\mathcal{T}_t$ designate a tree where there is a unique path from any node v_i to t (tree rooted at t as defined in [14]) and

π_{v_i} denote the cost of the path p, from v_i to t, in $\mathcal{T}_t$; the *reduced cost* $\bar{c}_{v_iv_j}$ of arc $(v_i,v_j)\in A$ associated with $\mathcal{T}_t$ is $\bar{c}_{v_iv_j}=\pi_{v_j}-\pi_{v_i}+c_{v_iv_j}$. So all arcs in $\mathcal{T}_t$ have a null reduced cost. The reduced cost of path p is given by $\sum_{(v_i,v_j)\in p}\bar{c}_{v_iv_j}$ and it can be proved that $c(p)=\bar{c}(p)+\pi_s$. The advantage of using reduced costs was first noted by Eppstein [16] and they are shown by Theorems 8 and 9 in [14] and by Theorem 2.1 in [15] (in the context of the MPS algorithm) to lead to less arithmetic operations and to sub-path generation simplification.

Let $\mathcal{T}_t^*$ be the tree of the shortest paths from all nodes to t and $\mathcal{T}_t^*(v_j)$ the shortest path from v_j to t in $\mathcal{T}_t^*$ (hence $\pi_{v_i}=c[\mathcal{T}_t^*(v_j)]$). The sub-path from v_k to v_j in p is represented by $\mathrm{sub}_p(v_k,v_j)$. The set of arcs of A of $G=(N,A)$ is arranged in the *sorted forward star form* – for details, see [17]. That is, the set A is sorted in such a way that, for any two arcs $(v_i,v_j),(v_k,v_l)\in A$, $(v_i,v_j)<(v_k,v_l)$ if $v_i<v_k$ or ($v_i=v_k$ and $\bar{c}_{v_iv_j}\leq\bar{c}_{v_kv_l}$).

3. Node Disjoint and SRLG Diverse Path Pairs

The algorithm is based on the Algorithm 1 in [12] for enumerating node disjoint path pairs, by non-increasing total additive cost which requires a network topology transformation as described in the next subsection.

3.1. Network Topology Modification

Let s,t be a source and destination in G. Let P_{xy} be the set paths (loopless or not) from node x to node y in G. Let $G'=(N',A')$ be a transformed network where, such that [12]:

- the former nodes are duplicated: $N'{=}N\cup\{v_i':v_i\in N\}$;
- the former arcs are duplicated, and a new one, linking t and the new node s', is added: $A'=A\cup\{a'=(v_a',v_b'):a=(v_a,v_b)\in A\}\cup\{(t,s')\}$;
- $c(v_a',v_b')=c(v_a,v_b),\quad\forall(v_a,v_b)\in A$;
- $c(t,s')=0$;
- $R_{a'}=R_a,\quad\forall a,a'\in A'$.

In this new network the source node is s and the destination node is t'. Each path from s to t' in G' is such that:

$$p=q\diamond(t,s')\diamond q', \tag{7}$$

where $q\in P_{st}$ and $q'\in P_{s't'}$. If q and q' are simple and do not share corresponding nodes in N and N' (except s, s' and t, t') then they are disjoint simple paths. If, additionally, $R_q\cap R_{q'}=\emptyset$, then q and q' are SRLG diverse.

Let $\mathcal{T}_{t'}^*$ be the tree of the shortest paths from all nodes to t', in G' (the modified graph). If $\mathcal{T}_t^*$ is calculated before transforming the network, then $\mathcal{T}_{t'}^*$ can easily be obtained. This process of building $\mathcal{T}_{t'}^*$ ensures that $\mathcal{T}_{t'}^*(s)=p\diamond(t,s')\diamond p'$, where p and p' correspond to the same path. In the transformed network, $\pi_{v_i'}=c(\mathcal{T}_t^*(v_i)),\forall v_i'\in N'\backslash N$ and $\pi_{v_i}=\pi_{v_i'}+\pi_{s'}$, for any $v_i\in N$ [12][1].

In Remark 1 of [12] it is suggested that there is no need to explicitly represent the new arcs in G' except the new arc (t,s'), because every new arc is a copy of another existing arc, and $\bar{c}_{v_i'v_j'}=\bar{c}_{v_iv_j}$. However, implementing Remark 1 is only feasible if $\mathcal{T}_{t'}^*$ is built as described in the previous paragraph – a fact which is not pointed out in [12].

If at least two different paths, p and q, with the same minimal cost exist from v_i to t, (with the successor of v_i in p different from the successor of v_i in q), then, using Dijkstra's algorithm in G' for calculating $\mathcal{T}_{t'}^*$, we may obtain $\mathcal{T}_{t'}^*(v_i)=(v_i,v_j)\diamond\mathcal{T}_{t'}^*(v_j)$ and $\mathcal{T}_{t'}^*(v_i')=(v_i',v_k')\diamond\mathcal{T}_{t'}^*(v_k')$, with $v_j\neq v_k$ (and $v_j'\neq v_k'$). When this happens, two different arcs with the "same tail", v_i and v_i', will belong to $\mathcal{T}_{t'}^*$, and when building the sorted forward star form of the arcs $A\cap(t,s')$, both arcs must be the first arc with tail v_i, which is not possible! This detail is very important because the MPS algorithm [14], which is the base of Algorithm 1 in [12], requires the ordering of the arcs in the ordered forward star form, such that the first arc with tail v_i (equivalent to v_i') $\forall v_i\in A$, belongs to $\mathcal{T}_{t'}^*$, in order to be able to generate every path by non-decreasing order of its cost.

3.2. The Algorithm

A infeasibility test can be made at the very beginning of the algorithm:

- if we can not find at least two arcs with tail node s, which are SRLG diverse, then there is no solution;
- if we can not find at least two arcs with head node t, which are SRLG diverse, then there is no solution.

If this infeasibility test fails, then we can proceed to try and find SRLG diverse path pairs.

In order to speed up path generation, the network should be pruned of the arcs with tail s and head t such that no SRLG diverse paths can be obtained if they belong to any of the paths. We will say that the remaining arcs of tail s and head t can be SRLG protected (by at least another arc of tail s or head t, respectively). These arcs can be identified during the infeasibility test and removed from the network[2] before running the Dijkstra algorithm for obtaining $T_{t'}$.

A path, p, obtained in the augmented network (see Subsection 3.1 or [12; Subsection 3.2]), is made of $q\diamond(t,s')\diamond q'$ and we assume it has deviation node d_p, deviation arc a_h,

[1] In [12], where is $\pi_i=\pi_{i'}+\pi_s$ should be $\pi_i=\pi_{i'}+\pi_{s'}$.

[2] In order to reduce the need for graph transformation, these arcs can be simply marked as useless, as long as an adequate Dijkstra's algorithm is implemented.

```
Algorithm 1: Determination of the K shortest SRLG diverse simple path pairs

Data: Network directed graph $G = (N,A)$ and a source destination node pair $(s,t)$, and $c$ cost of the links
Result: $S$, the set of the $K$ shortest SRLG diverse simple path pairs from $s$ to $t$
1  if the infeasibility test is successfully then Stop end
2  Remove from $A$ arcs emerging form $s$ or incident in $t$, which can not be SRLG protected. Remove from $A$ all arcs with tail node $t$
3  $\mathcal{T}^*_{t'} \leftarrow$ tree of the shortest paths from $i \in N'$ to $t'$ using $c$
4  $p \leftarrow \mathcal{T}^*_{t'}(s)$
5  if $p$ is not defined then Stop end
6  $\bar{c}_{v_i vj} \leftarrow \pi_{v_j} - \pi_{v_i} + c_{v_i v_j}, \quad \forall (v_i, v_j) \in A'$
7  Represent $A'$ in the sorted forward star form concerning $\bar{c}$
   // Consider: $p = (s \equiv v_1, v_2, ..., v_{y-1}, v_y \equiv t)$, $(s, v_2)$ can be SRLG protected
8  $d_p \leftarrow s$ // Deviation node of $p$
9  $X \leftarrow \{p\}$
10 $S = \emptyset$
11 while $X \neq \emptyset \wedge |S| < K$ do
12     $p \leftarrow$ path in $X$ such that $\bar{c}(p)$ is minimum
13     if ($p$ is simple) $\wedge$ Disjoint($p$) $\wedge$ SRLGDiverse($p$) then
14         $S \leftarrow S \cup \{p\}$
15     end
16     $X \leftarrow X \backslash \{p\}$
17     $i \leftarrow$ min index such that $v_i = d_p$
18     break $\leftarrow$ false // Candidate paths might be derived from $p$
19     repeat
20         $l \leftarrow$ index such that $a_l = (v_i, v_{i+1})$
21         repeat
22             $l \leftarrow l + 1$
23             $v_j \leftarrow$ head node of $a_l$ // if $l > m + 1$ then $v_j \leftarrow 0$
24             if ($v_i$ is the tail node of $a_l$)$\wedge$ EquivalentPair(sub$_p(s, v_i) \diamond a_l \diamond \mathcal{T}^*_{t'}(v_j)$) then
25                 break $\leftarrow$ true // No candidate paths will derive from $p$ at $v_i$
26             end
27         until break $\vee$($v_i$ is not the tail node of $a_l$)$\vee$ [($a_l$ does not form a loop with sub$_p(s, v_i)$)$\wedge$ SRLGDiverse(sub$_p(s, v_i) \diamond a_l$) $\wedge$ Disjoint(sub$_p(s, v_i) \diamond a_l$)]
28         if ($\neg$ break )$\wedge$($v_i$ is the tail node of $a_l$) then
29             $q \leftarrow$ sub$_p(s, v_i) \diamond a_l \diamond \mathcal{T}^*_{t'}(v_j)$; $d_q \leftarrow v_i$
30             $X \leftarrow X \cup \{q\}$
31         end
32         $v_i \leftarrow v_{i+1}$ // Next node of $p$
33     until ($v_i = t'$) $\vee$ $\neg$(sub$_p(s, v_i)$ is simple) $\vee$ $\neg$ Disjoint(sub$_p(s, v_i)$) $\vee$ $\neg$ SRLGDiverse(sub$_p(s, v_i)$)
34 end
```

and that the first arc in q is a_f (where $a_f = a_h$ if the deviation node is s). Paths will only be placed in the set of candidate paths if:

- the deviation node, d_p, belongs to N and the path sub$_p(s, d_p) \diamond a_h$ is simple;
- the deviation node, d_p, belongs to $N' \backslash N$:
  - the path sub$_p(s', d_p) \diamond a_h$ is simple;
  - $c(\text{sub}_p(s,t)) \geq c(\text{sub}_p(s',t'))$ (or $c(q) \geq c(q')$); note that $c(q') = c(\text{sub}_p(s', d_p) \diamond a_h)$, and that $c(q') = c(p) - c(q)$;
  - the paths sub$_p(s,t)$ and sub$_p(s', d_p) \diamond (a_h)$ are node-disjoint;
  - a_h is SRLG diverse with sub$_p(s,t)$.

In Algorithm 1 we chose to remove from the network graph arcs which are not useful for obtaining SRLG diverse path pairs. This is not strictly necessary, but improves the algorithm efficiency.

Note that in set X all paths are simple, disjoint and SRLG diverse up to and including the deviation arc. Due to this fact we have replaced all the interior **while** cycles of Algorithm 1 [12] with **repeat until** cycles.

Function **Disjoint**(p), $p = q \diamond (t, s') \diamond q'$, returns true if q and q' are node disjoint. Function **SRLGDiverse**(p) returns true if q and q' are SRLG diverse. At Steps 27 and 33 the value of functions **Disjoint()** and **SRLGDiverse()** is true whenever v_i belongs to N. This implies that the evaluation of disjointness or SRLG diverseness is only effectively required at Steps 27 and 33 of the algorithm when the deviation node belongs to $N' \backslash N$. Also note that for the calculation of **SRLGDiverse**(sub$_p(s, v_i) \diamond a_l$), in Step 27, it is sufficient to evaluate if sub$_p(s, v_i)$ is SRLG diverse with arc a_l.

Function **EquivalentPair()** was first introduced in [12], for including Remark 2 in Algorithm 1. Due to Remark 2 in [12] we may choose to store paths pairs that $\bar{c}[\text{sub}_p(s,t)] \leq \bar{c}[\text{sub}_p(s',t')]$ or $\bar{c}[\text{sub}_p(s,t)] \geq \bar{c}[\text{sub}_p(s',t')]$. If we choose to store in X paths q such that $\bar{c}[\text{sub}_q(s,t)] \geq \bar{c}[\text{sub}_q(s',t')]$, then function **EquivalentPair()** will only be required when v_i belongs to $N' \backslash N$ – that is Step 24 could be rewritten:

$(v_i \in N' \backslash N) \wedge (v_i$ is the tail node of $a_l) \wedge$
EquivalentPair(sub$_p(s, v_i) \diamond a_l \diamond \mathcal{T}^*_{t'}$(head node of a_l)).

Function **EquivalentPair**(p) returns true whenever $\bar{c}[\text{sub}_p(s,t)] < \bar{c}[\text{sub}_p(s',t')]$. Consider that v_i belongs to $N' \backslash N$ and let $q = \text{sub}_p(s, v_i) \diamond a_l \diamond \mathcal{T}^*_{t'}$(head node of a_l), in Step 24. In this case sub$_p(s,t) = $ sub$_q(s,t)$, therefore the execution of **EquivalentPair()** can be simply the evaluation of $\underbrace{\bar{c}[\text{sub}_p(s,t)]}_{\bar{c}[\text{sub}_q(s,t)]} < \underbrace{\bar{c}(q) - \bar{c}[\text{sub}_p(s,t)]}_{\bar{c}[\text{sub}_q(s',t')]}$.

The proposed algorithm requires a directed network graph. For obtaining SRLG diverse path pairs in an undirected network, we must build the equivalent directed graph: each undirected link is represented by two directed arcs, in opposite directions,with the same costs, belonging to the same SRLGs as the corresponding undirected link.

3.3. Link Disjoint SRLG Diverse Path Pairs

If the path pair does not need to be node disjoint, then the only modification required in Algorithm 1 is the suppression of the function **Disjoint()**, assuming each undirected link belongs to at least one SRLG.

3.4. SRLG Diverse Path Pairs With Length Constraints

Let $p = q \diamond (t,s') \diamond q'$, represent a path pair (q,q'). If the path pairs have length restrictions (maximum number of allowed arcs), then two new conditions must be evaluated: the depth of the deviation node $i \in q$ and $j' \in q'$ must be less than the length constraint (assuming node s has depth 0).

4. Computational Results

Two sets of experiments were made. The first set used:

- Randomly generated undirected networks with $n = 25, 50, 100, 200, 400$ and $m = 3n, 4n$ (where n is the number of nodes and m is the number of undirected arcs).
- The cost of each link was randomly generated in [1..65535].
- Each undirected arc was associated with a single SRLG.
- For each value of n and m ten randomly generated networks were considered.
- For each network 50 end-to-end node pairs, where selected and $K = 1000$ diverse path pairs were sought.

The second set of experiments considered:

- Randomly generated undirected networks with $n = 100, 1000$ and $m = 3n, 4n$ (where n is the number of nodes and m is the number of undirected arcs).
- The cost of each link was randomly generated in [1..65535].
- Each undirected arc was associated with a single SRLG.
- For each value of n and m ten randomly generated networks were considered.
- For each network 50 end-to-end node pairs, where selected and $K = 5000$ diverse path pairs were sought.

The computer used was a PC, Intel(R) Core(TM) 2 Duo Processor, 1.82 GHz, RAM 1 GB, under Kubuntu. The maximum number of allowed paths was 10^7.

Observing the average central processing unit (CPU) time per node pair, for obtaining $K = 1000$ path pairs, presented in Figs. 1 and 2, it may be concluded that it is more efficient for obtaining node disjoint than arc disjoint path pairs, as expected.

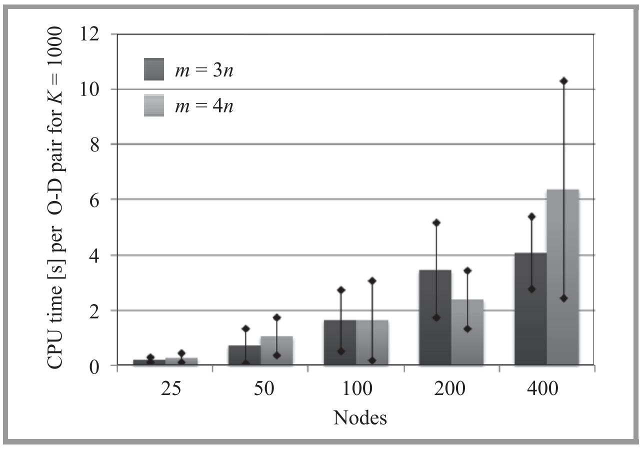


Fig. 1. CPU times for obtaining $K = 1000$ arc disjoint and SRLG diverse path pairs.

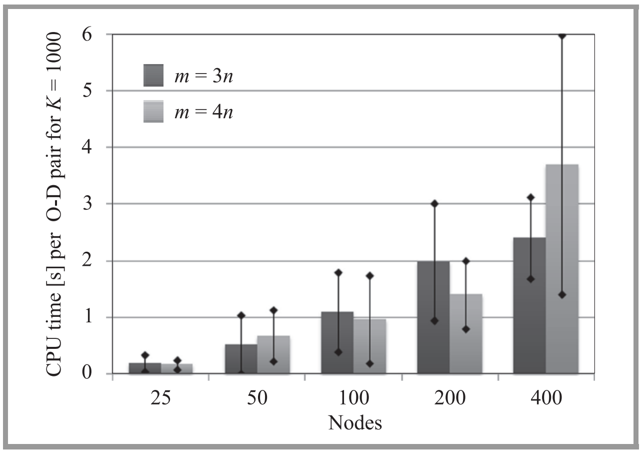


Fig. 2. CPU times for obtaining $K = 1000$ node disjoint and SRLG diverse path pairs.

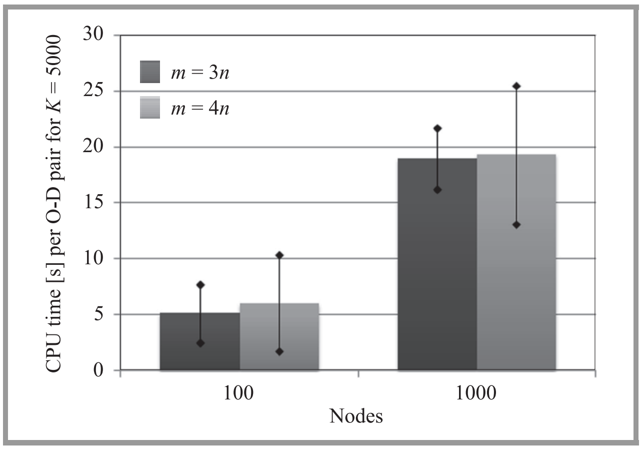


Fig. 3. CPU times for obtaining $K = 5000$ arc disjoint and SRLG diverse path pairs.

That statement is still true if $K = 500$ is used, as can be seen in Figs. 3 and 4. It should be noted that with $n = 100$ when K goes from 1000 to 5000, the CPU time grows proximately linearly with K. Also the CPU time with $n = 1000$, in Figs. 3 and 4, is less than 10 times the CPU time when $n = 100$.

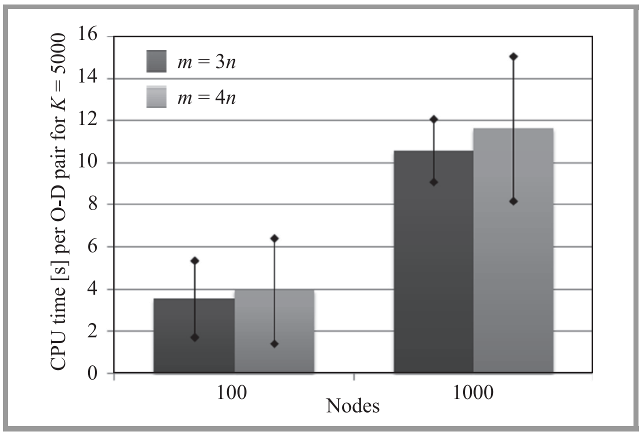


Fig. 4. CPU times for obtaining $K = 5000$ node disjoint and SRLG diverse path pairs.

The interval bars in the figures indicate the 95% confidence interval for the average CPU time.

The CPU time grows with the number of nodes, for the networks with the same average degree, and also tends to increase with the average node degree, for networks with the same number of nodes. In the cases where that does not happen, it was due to some(s) node pair(s) with a CPU quite above the average CPU time in one (or two) of the ten networks. The average CPU times presented in Figs. 1–4 were obtained for the node pairs for which the desired K (1000 or 5000) were obtained.

The CPU time is closely related to the total number of candidate paths that were generated and added to set X in the algorithm, as it can be seen in Figs. 5–8.

In some cases there is no solution (and in most cases this was discovered very fast due to the infeasibility test) or the maximum allowed number (10^7) of candidate path pairs was generated without obtaining the desired number K of SRLG disjoint path pairs – in this last case it is uncertain whether any more disjoint path pairs might have been obtained if the maximum allowed number of candidate paths was higher. In the experiments, we tried to obtain $K = 1000$ SRLG node disjoint path pairs for 5000 node pairs (5 different values of n, two different average node degrees, 10 different seeds for network generation and 50 node pairs per network) and failed to do it in 22 cases due to the fact that the maximum number of candidate paths was attained. This corresponds to a success rate of 99.56%. The results for arc and SRLG disjoint path pairs was similar: 99.54%. When $K = 1000$ the unsuccessful node pairs occurred only for $n = 200$ and $n = 400$ and the CPU times is approximately 2 minutes and 4 minutes, for the node and arc disjoint path pairs, respectively.

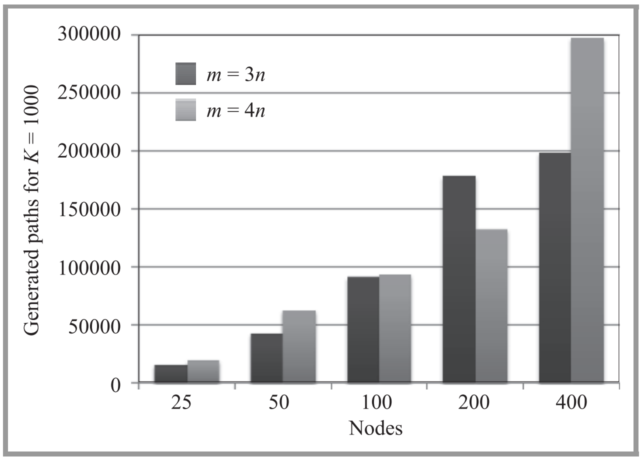


Fig. 5. Number of candidate path pairs added to X for obtaining $K = 1000$ arc disjoint and SRLG diverse path pairs.

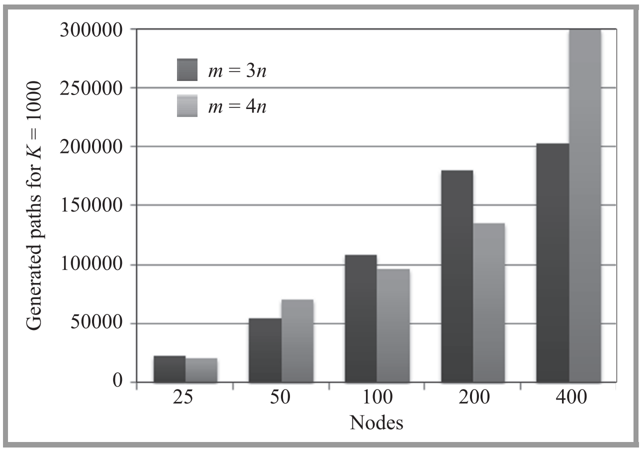


Fig. 6. Number of candidate path pairs added to X for obtaining $K = 1000$ node disjoint and SRLG diverse path pairs.

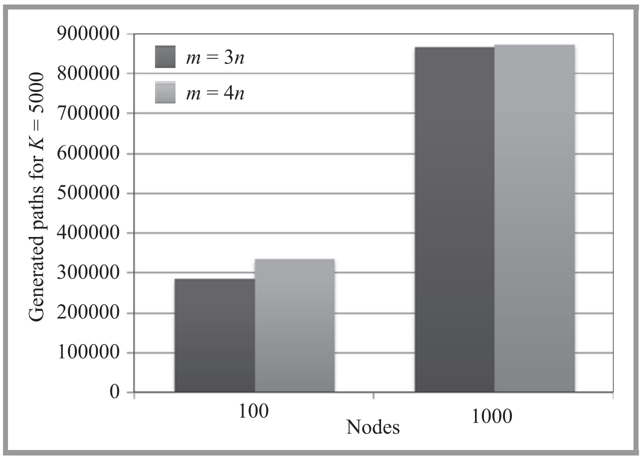


Fig. 7. Number of candidate path pairs added to X for obtaining $K = 5000$ arc disjoint and SRLG diverse path pairs.

When $K = 5000$ the rate of unsuccessful node pairs grows to 7.4% for $n = 1000$ but the CPU time remains similar to what was observed, for the unsuccessful cases when $K = 1000$.

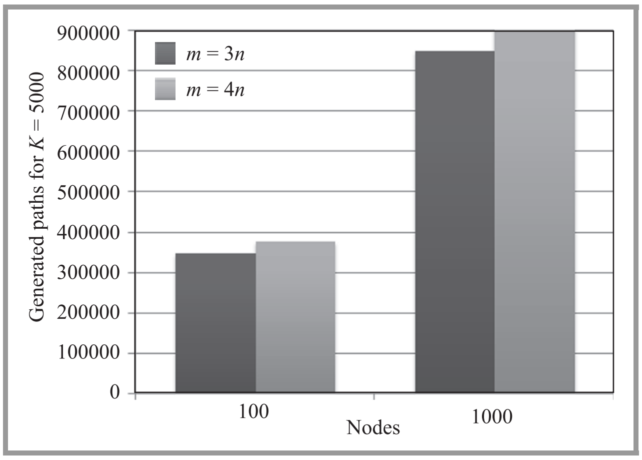


Fig. 8. Number of candidate path pairs added to X for obtaining $K = 5000$ node disjoint and SRLG diverse path pairs.

Note that in the case $n = 25, 50, 100, 200, 400$ the algorithm was able to detect that no solution could be found for thirteen node pairs in zero seconds. However, for $n = 1000$ there were two node pairs for which no solution was found even after generating 10^7 candidate path pairs, as it can be seen in Tables 1 and 2, in the line with "$k = 0$ (?)".

Table 1
Arc disjoint and SRLG diverse path pairs, for $m = 3n, 4n$: total number of node pairs and average CPU times when $K = 5000$ was not obtained

Arc disjoint	$m = 3n$		$m = 4n$	
n	100	1000	100	1000
$0 < k < 5000$	3	34	2	31
CPU(s)	212	244	222	232
$k = 0$ (?)	–	1	–	–
CPU(s)	–	277	–	–

Table 2
Node disjoint and SRLG diverse path pairs, for $m = 3n, 4n$: total number of node pairs and average CPU times when $K = 5000$ was not obtained

Node disjoint	$m = 3n$		$m = 4n$	
n	100	1000	100	1000
$0 < k < 5000$	4	36	1	29
CPU(s)	125	139	125	132
$k = 0$ (?)	–	1	–	–
CPU(s)	–	155	–	–

The results show that the algorithm solves exactly the problem of obtaining K path pairs, node disjoint (arc disjoint) and SRLG diverse in most cases. When the algorithm generates, k, $0 \leq k < K$ paths (due to the allowed maximum number of generated paths), this can be CPU time consuming. Therefore, in order to avoid the high CPU time, sometimes required by the algorithm, a CPU time limit should be imposed for obtaining the desired number of solutions (K), so that it can be used as a subroutine in a multicriteria approach to reliable routing taking into account SRLGs.

5. Conclusion

The multi-layer nature of telecommunication networks, makes it more difficult to implement recovery mechanisms to ensure routing resiliency. The introduction of the concept of SRLGs allows an upper layer to select, for a given active path, a backup path, which avoids every SRLG that may involve the selected AP in the event of afailure. A formulation of the SRLG diverse path pair calculation problem, in a directed network was put forward. An exact algorithm for enumerating SRLG diverse paths in (un)directed networks, by non-decreasing cost of their total (additive) cost was proposed. The considered SRLG diverse path pairs may be node or arc disjoint, with or without length constraints.

Computational results displayed the efficiency of the proposed algorithm for obtaining node or arc disjoint SRLG diverse path pairs in undirected networks. The experimental results show the algorithm solves the problem of enumerating $K = 1000$ disjoint paths pairs in most cases, using less than one second for the smaller networks. However, when the desired value for K can not be attained, the CPU time can grow significantly. Considering that this algorithm can be used as a subroutine in a multicriteria approach to resilient routing, taking into account SRLGs, we would advise a lower number of allowed maximum candidate paths or a limit of CPU time per node pair.

Acknowledgements

We thank Marco António do Carmo Silva for collecting the data and producing the figures used in this work.

This work was partially supported by programme POSC of the EC programme cosponsored by national funds.

References

[1] J. Q. Hu, "Diverse routing in optical mesh networks", *IEEE Trans. Commun.*, vol. 51, no. 3, pp. 489–494, 2003.

[2] M. J. Rostami, S. Khorsandi, and A. A. Khodaparast, "CoSE: ASRLG-disjoint routing algorithm", in *Proc. Fourth Eur. Conf. Univ. Multiserv. Netw. ECUMN'07*, Toulouse, France, 2007.

[3] D. Xu, Y. Chen, Y. Xiong, C. Qiao, and X. He, "On finding disjoint paths in single and dual link cost networks", in *IEEE INFOCOM 2004 Conf.*, Hong Kong, 2004.

[4] A. K. Todimala and B. Ramamurthy, "IMSA: an algorithm for SRLG diverse routing in WDM mesh networks", in *Proc. Int. Conf. Comput. Commun. Netw. ICCCN 2004*, Chicago, USA, 2004, pp. 199–204.

[5] J. W. Suurballe and R. E. Tarjan, "A quick method for finding shortest pairs of disjoint paths", *Networks*, vol. 14, no. 2, pp. 325–336, 1984.

[6] R. Bhandari, *Survivable Networks, Algorithms for Diverse Routing.* Norwell: Kluwer, 1999.

[7] A. K. Todimala and B. Ramamurthy, "A heuristic with bounded guarantee to compute diverse paths under shared protection in WDM mesh networks", in *Proc. IEEE Globlecom 2005 Conf.*, St. Louis, USA, 2005, pp. 1915–1919.

[8] P. Laborczi, J. Tapolcai, P.-H. Ho, T. Cinkler, A. Recski, and H. T. Mouftah, "Algorithms for asymmetrically weighted pair of disjoint paths in survivable networks", in *Proc. Des. Rel. Commun. Netw. DCRN 2001*, Budapest, Hungary, 2001, pp. 220–227.

[9] X. Pan and G. Xiao, "Algorithms for the diverse routing problem in WDM networks with shared risk link groups", in *Int. Conf. Comput. Sci. ICCS 2004*, Krakow, Poland, 2004, pp. 381–385.

[10] X. Pan and G. Xiao, "Heuristics for diverse routing in wavelength-routed networks with shared risk link groups", *Phot. Netw. Commun.*, vol. 11, no. 1, pp. 29–38, 2006.

[11] D. Xu, Y. Xiong, C. Qiao, and G. Li, "Trap avoidance and protection schemes in networks with shared risk link groups", *J. Lightw. Technol.*, vol. 21, no. 11, pp. 2683–2693, 2003.

[12] J. C. N. Clímaco and M. M. B. Pascoal, "Finding non-dominated bicriteria shortest pairs of disjoint simple paths", *Comput. Oper. Res.*, vol. 36, no. 11, pp. 2892–2898, 2009.

[13] J. Bang-Jensen and G. Gutin, *Digraphs: Theory, Algorithms and Applications*, Springer Monographs in Mathematics. Great Britain: Springer, 2002.

[14] E. Martins, M. Pascoal, and J. Santos, "Deviation algorithms for ranking shortest paths", *Int. J. Found. Comput. Sci.*, vol. 10, no. 3, pp. 247–263, 1999.

[15] E. Martins, M. Pascoal, and J. Santos, "An algorithm for ranking loopless paths", Tech. Rep. 99/007, CISUC, 1999 [Online]. Available: http://www.mat.uc.pt/∼marta/Publicacoes/mps2.ps

[16] D. Eppstein, "Finding the k shortest paths", *SIAM J. Comput.*, vol. 28, no. 2, pp. 652–673, 1999.

[17] R. Dial, F. Glover, D. Karney, and D. Klingman, "A computational analysis of alternative algorithms and labeling techniques for finding shortest path trees", *Networks*, vol. 9, no. 3, pp. 215–348, 1979.

Teresa Gomes is Assistant Professor in telecommunications at the Department of Electrical Engineering and Computers of the Faculty of Sciences and Technology of the University of Coimbra, Portugal, since 1998, and a researcher at the INESC-Coimbra. She obtained the following degrees: undergraduate diploma in electrical engineering science (E.E.S.)-informatics at the Coimbra University (1984), M.Sc. in computer science (1989) and Ph.D. in E.E.S.-telecommunications and electronics (1998), both at the University of Coimbra. Her main present interests are routing, protection and reliability analysis models and algorithms for optical and MPLS networks.

e-mail: teresa@deec.uc.pt
Department of Electrical Engineering and Computers
University of Coimbra
Pinhal de Marrocos
3030-290 Coimbra, Portugal

José Craveirinha is Full Professor in telecommunications at the Department of Electrical Engineering and Computers of the Faculty of Sciences and Technology of the University of Coimbra, Portugal, since 1997. He obtained the following degrees: undergraduate diploma in electrical engineering science (E.E.S.)-telecommunications and electronics at IST, Lisbon Technical University (1975), M.Sc. (1981) and Ph.D. in E.E.S. at the University of Essex (UK) (1984) and Doct. of Science (Agregado) in E.E.S.-telecommunications at the University of Coimbra (1996). Previous positions were: Associate Professor and Assistant Professor at the FCUT, Coimbra University, telecommunication R&D engineer (at the CET-Portugal Telecom). He coordinated a research group in teletraffic engineering and network planning at the INESC-Coimbra R&D Institute since 1986 and was Director of this Institute during 1994–1999. He is author/co-author of more than 100 scientific and technical publications in teletraffic modeling, reliability analysis, planning and optimization of telecommunication networks. His main present interests are in multicriteria routing and reliability analysis models and algorithms for optical and multiservice-IP/MPLS networks.

e-mail: jcrav@deec.uc.pt
Department of Electrical Engineering and Computers
University of Coimbra
Pinhal de Marrocos
3030-290 Coimbra, Portugal

Paper

Performance Analysis of a Bi-Objective Model for Routing and Wavelength Assignment in WDM Networks

Carlos Simões, Teresa Gomes, José Craveirinha, and João Clímaco

Abstract—Establishing end-to-end connections on wavelength division multiplexing (WDM) networks requires setting up lightpaths, defining the sequence of optical fibres and the wavelength in each fibre (the routing and wavelength assignment problem) for traffic flow. This paper reviews a bicriteria model for obtaining a topological path (unidirectional or symmetric bidirectional) for each lightpath request in a WDM network, developed by the authors, and presents a performance analysis of the model by considering important network performance measures. An extensive performance analysis of the two bicriteria model is presented, comparing the performance metrics obtained with the monocriterion models using the same objective functions, in five different reference networks commonly used in literature.

Keywords—multicriteria optimization, routing in WDM networks.

1. Introduction

1.1. Background Concepts

All-optical networks based on wavelength division multiplexing (WDM) have emerged as a promising technology for network operators to respond to an increased demand for broadband services, exploiting the huge bandwidth of optical fibres. All-optical networks based on wavelength division multiplexing consist of optical fibre links and nodes, and the WDM scheme divides the optical bandwidth into independent channels, each one with a different wavelength, operating at transmission rates compatible with the lower capacity of the end user's devices. Each node in a all-optical network has a dynamically configurable optical switch or router which supports wavelength based switching or routing. Configuring these optical devices across the network enables that node pairs can establish point-to-point all-optical connections, or lightpaths, for information transfer. A lightpath may span several fibre links and consist of wavelength channels in the sequence of this links, interconnected at the nodes by means of optical routing. In order to establish a lightpath, the network needs to decide on the topological route and the wavelength(s) for the lightpath. In the absence of wavelength converters, a lightpath must use the same wavelength on all the links of its route (the *wavelength continuity constraint*), but wavelengths can be reused by different lightpaths in the network, as long as they do not share any fibre link.

Given a set of connection requests, the problem of setting up lightpaths by defining a path and assigning a wavelength to each of its links for every connection is called the routing and wavelength assignment (RWA) problem.

If the network nodes have wavelength converters, it is possible to assign different wavelengths on the multiple links of the lightpath. As a result, the wavelength continuity constraint is relaxed, thereby increasing the possible number of lightpaths that can simultaneously be established in the network. However, since wavelength converters are costly and may cause signal quality degradation, often no wavelength converters are used or only some nodes have this capability. The converter configuration of the network is called full if all nodes have wavelength converters and sparse if only a part of the nodes have them. Obviously, wavelength conversion leads to lower blocking probabilities, but, in practice, some works have shown that with only a small number of converters placed in strategic locations, a significant performance improvement can be achieved [1]. On the other hand, when a node is capable of converting a wavelength to any other wavelength, the node is said to have complete conversion capability. If a node is able to convert an incoming wavelength to only a subset of available wavelengths, the node is said to have limited or partial conversion capability. A wavelength converter is said to have a conversion degree D, if it can shift any wavelength to one of D wavelengths.

Multi-fibre networks use several fibres per link. Considering that the nodes have no wavelength converters, the possibilities of finding a lightpath satisfying the wavelength continuity constraint is higher than in single fibre networks. A multi-fibre network with F fibres per link and W wavelengths per fibre is functionally equivalent to a single-fibre network with $F \times W$ wavelengths and conversion degree of F [2].

Connection requests are usually considered to be of three types: static, incremental, and dynamic [3]. Regarding static traffic, the entire set of connections is known in advance, and the problem is then to set up lightpaths for these connections seeking to minimize network resources such as the number of wavelengths or the number of fibres in the network. In this case the RWA problem is known as the static lightpath establishment (SLE) problem. Concerning incremental-traffic, connection requests arrive sequentially, a lightpath is established for each connection, and the lightpath remains in the network indefinitely. In the case of dynamic traffic, a lightpath is set up for each connection request as it arrives, and the lightpath is released after some finite amount of time. The objective in the incremental and dynamic traffic cases is to set up lightpaths and assign wavelengths seeking to minimize the amount of connection blocking, or to maximize the number of connections that are established in the network at any time. This problem is designated as the dynamic lightpath establishment (DLE) problem. The SLE problem can be formulated as an integer linear program [4], which is known to be NP-complete [5].

In order to make the RWA problem more tractable, it can be divided into two sub-problems – routing and wavelength assignment, so that each sub-problem can be solved separately. Nevertheless note that each sub-problem is still NP-complete [5].

Routing methods in WDM networks can be classified into two types: static routing and adaptive routing. Static routing encompasses fixed routing and fixed-alternate routing. In fixed routing the same fixed route for a given source-destination pair remains unchanged throughout time. The fixed-alternate routing scheme pre-computes a set of paths between each source-destination pair, and for each request, a path from this pre-computed set is chosen. Adaptive routing involves a dynamical search for a path when a connection request arrives, taking into account the current state of the network. Therefore in general, adaptive routing gives better blocking performance than fixed-alternate routing [3].

A wavelength assignment algorithm is used to determine the wavelengths in the arcs along the path chosen in the routing step. Many wavelength assignment algorithms have been proposed such as random, first fit, most-used, least-used, least-loaded, max-sum, min-product, and relative capacity loss schemes [3].

In most approaches presented in the literature, routing and wavelength assignment are *optimized* separately by considering a decomposition of the global RWA problem through heuristic algorithms, because these problems are NP-complete. However, some algorithms consider the routing and the wavelength assignment jointly [6], [7].

Many different integer linear programming (ILP) formulations have been proposed for the RWA problem in WDM optical networks, under different objectives. However, although those formulations lead to exact solutions, most of the times they have not been used for developing solution schemes except for very small networks or for some rounding off procedures [8]. Examples of monocriterion approaches to the RWA problem, considering different metrics as objective functions are in the references: [3], [4], [6], [7], [9]–[12].

1.2. Multicriteria Models

In general routing protocols only optimize one metric, typically using some variant of a shortest path algorithm. However all-optical WDM networks can be characterized in terms of performance by multiple metrics. Also the design of real networks usually involves multiple, often conflicting objectives and various constraints. The development of multicriteria models that explicitly represent the different performance objectives, enabling to treat in a consistent manner the trade off among the various criteria, seems to be potentially advantageous in face of the inherent limitations of single objective approaches.

Note that in models involving explicitly multiple and conflicting criteria, the concept of optimal solution is replaced by the concept of non-dominated solutions. A non-dominated solution is a feasible solution such that no improvement in any criterion may be achieved without sacrificing at least one of the other criteria.

A state-of-art review on multicriteria approaches in communication networks was presented in [13], including a section dedicated to routing models. A recent review on multicriteria routing models can be seen in [14].

In [15] two different criteria, path length and congestion in the network, are considered and applied sequentially (the second metric is only used if a tie occurs in the first one). Two algorithms for dynamic traffic were proposed: least congested shortest hop routing where priority is given to efficient resource utilization (the algorithm selects the least congested path among all shortest hop paths currently available); and shortest hop least congested routing in which priority is given to maintaining the load balance in the network (it selects the shortest hop path among all the least congested paths). This type of models may be considered as a first-tentative multicriteria approach as analyzed in [13].

In [16] a set of link disjoint routes for each node pair, in a network with dynamic traffic, is pre-computed. Then the weighted least-congestion routing and first-fit wavelength assignment algorithm, which includes two criteria (hop count and free wavelengths), combined in a single weighted metric, is used to rank the paths. The same approach was proposed earlier in [17], which tries to minimize the resource utilization while simultaneously balancing the traffic load in the links. Nonetheless [17] only considers networks with full wavelength conversion. These models, which consider as solution of the bicriteria problem the optimal solution of the single objective function resulting from the weighted sum of the considered criteria, do not take full advantage of the possibilities of multicriteria approaches and may lead to less effective solutions.

A bicriteria model for obtaining a topological path (unidirectional or symmetric bidirectional) for each lightpath request in a WDM network with multi-fibre links and an exact resolution approach for that model was presented by the authors in [18]. The first criterion is related to bandwidth usage in the links (or arcs) of the network. The second criterion is the number of links (hops) of the path. The resolution approach [18] uses an exact procedure to calculate non-dominated topological paths based on a *k*-shortest path algorithm [19] which is based on an adaptation of the MPS algorithm [20]. Furthermore, preference thresholds, defined in the objective function's space, combined with a Chebyshev distance to a reference point [21] are used for selecting the final solution. The solution of this bicriteria model is a non-dominated topological (optically feasible) path. A heuristic procedure is then used to assign wavelengths to the links.

The focus of this paper is to present an extensive and systematic performance analysis study of the bicriteria routing model [18] with respect to certain network performance measures by comparison of their results with the results of the associated single objective models, one related to the bandwidth usage and another that minimizes the number of used links (hop count). An incremental traffic model (where the duration of the connections is assumed unlimited) will be considered in several benchmark networks used in previous works in the area of WDM networks. The selected network performance measures are: the frequency of rejected requests (global blocking probability estimate), total used bandwidth, mean hop count of accepted requests, percentage of links with minimal free bandwidth, the average CPU time per request, and the percentage of non-optimal solutions.

The paper is organized as follows. In Section 2 the model without protection is described, together with the resolution approach of the bicriteria model. Performance analysis of the results obtained using several network topologies are presented and discussed in Section 3. Finally, some conclusions are drawn in Section 4.

2. The Bicriteria Routing Model

2.1. Model Description

In this section we review the features of the bicriteria routing model associated with the dynamic lightpath establishment problem with incremental traffic, in a WDM network, proposed in [18]. The model was developed for application in large WDM networks, with multiple wavelengths per fibre and multi-fibres per link. The bicriteria routing model considers the DLE problem with incremental traffic, and a mixture of unidirectional and bidirectional (symmetric) connections. In order to cover a wide variety of networks, different types of nodes are considered (with complete wavelength conversion capability, limited range conversion or no wavelength conversion capability) in the model. Due to the real-time nature of the intended application, solutions should be obtained in a short time. This requirement lead to the separation of the routing and wavelength assignment problems, having in mind an automatic selection of the solution (among the non-dominated solutions, previously identified). The wavelength assignment problem is solved separately, after the bicriteria routing problem.

Let $R = \{N, L, C, T_N\}$ represent the WDM network where:

- N is the set of nodes, $N = \{v_1, v_2, \ldots, v_n\}$, $n = \#N$.
- L is the set of directed arcs, $L = \{l_1, l_2, \ldots, l_m\}$, $m = \#L$.
- Set of wavelengths, $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_W\}$, $W = \#\Lambda$.
- Set of fibres, $F = \{f_1, f_2, \ldots, f_k\}$, $k = \#F$.
- Let $l_i = (v_a, v_b, \bar{o}_{l_i})$, $\bar{o}_{l_i} = (o_{l_i 1}, o_{l_i 2}, \ldots, o_{l_i k})$, $v_a, v_b \in N$.

  If $o_{l_i j} = (1, \bar{a}_j)(j = 1, 2, \ldots, k)$, then fibre f_j belongs to arc l_i and contains the wavelengths signalled in $\bar{a}_j$, $\bar{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jW})$ where $a_{ju} = 0, 1, 2$ $(u = 1, 2, \ldots, W)$:

  $$a_{ju} = \begin{cases} 0, \text{ if } \lambda_u \text{ does not exist in fibre } f_j \\ 1, \text{ if } \lambda_u \text{ exists and is free in fibre } f_j \\ 2, \text{ if } \lambda_u \text{ exists but is busy in fibre } f_j \end{cases} . \quad (1)$$

  If $o_{l_i j} = (0, \bar{a}_j)$ $(j = 1, 2, \ldots, k)$, fibre f_j does not belong to arc l_i.
- C is the arc capacity, $C(l_i) = (\bar{n}_{l_i}, \bar{b}_{l_i})$, with $\bar{n}_{l_i} = (n_{l_i 1}, n_{l_i 2}, \ldots, n_{l_i W})$ and $\bar{b}_{li} = (b_{l_i 1}, b_{l_i 2}, \ldots, b_{l_i W})$ where $n_{l_i j}$ is the total number of fibres in arc l_i with wavelength λ_j and $b_{l_i j}$ is the number of fibres where that wavelength is free in arc l_i.
- $T_N(v_i)$ is a table for each node $v_i \in N$ which represents the wavelength conversion capability of the nodes, that is the possibility of transferring the optical signal from one input λ_i to an output λ_j in the node:

  $$T_N(v_i) = [t_{uv}], \quad \forall v_i \in N; u, v = 1, 2, \ldots, W, \quad (2)$$

  where $t_{uv} = 1(0)$ whether (or not) λ_u can be converted into λ_v, in node v_i.

A *topological path*, p in R, is described by: a source node, a destination node $(v_s, v_t \in N)$ and the ordered sequence of nodes and arcs in the path, $p = \langle v_1, l_1, v_2, \ldots, v_{i-1}, l_{i-1}, v_i \rangle$, such that the tail of arc l_k is v_k and the head of l_k is v_{k+1}, for $k = 1, 2, \ldots, i-1$ (all the v_i in p are different).

Besides the ordered sequence of nodes and arcs, a *lightpath* p^λ also comprises the fibre used in each arc and the wavelength on the fibres:

$$p^\lambda = \langle l_c^*, \ldots, l_d^* \rangle = \langle (v_s, v_u, f_i, \lambda_\alpha), \ldots, (v_x, v_t, f_j, \lambda_\beta) \rangle \quad (3)$$

where $f_i, \ldots, f_j \in F$, $\lambda_\alpha, \ldots, \lambda_\beta \in \Lambda$, represent fibres and wavelengths, respectively.

Note that l_c^* corresponds to $l_c = (v_s, v_u, \bar{o}_{l_c})$ which implies $o_{l_c i} = (1, \bar{a}_i)$ and if $a_{i\alpha} = 1$ then $a_{i\alpha}$ will change from 1 to 2 if p^λ is selected.

A *bidirectional lightpath* $p^\lambda = (p_{st}^\lambda, p_{ts}^\lambda)$ is supported by a *bidirectional topological path* $p = (p_{st}, p_{ts})$, which is a pair of symmetrical topological paths.

Firstly we will describe the bicriteria model used for calculating topological paths.

The first objective function, $c_1(p)$ is related to the bandwidth usage in the links of the path p and is expressed in the inverse of the available bandwidth in the links:

$$\min_{p\in D}\left\{c_1(p)=\sum_{l\in p}\frac{1}{b_l^T}\right\}, \tag{4}$$

where D is the set of topological paths for the origin–destination node pair and b_l^T is the total available capacity in link l, in terms of available wavelengths. This criterion seeks to avoid already congested links, favoring a balanced distribution of traffic throughout the network, and hence decreasing the blocking probability and therefore increased the expected revenue.

Note that the values of the available bandwidths b_l^T to be used in each instance of the resolution of the bi-objective optimization problem are directly calculated from the vector $\bar{b}_l$ in $C(l)$:

$$b_l^T=\sum_{j=1}^{W} b_{lj}, \quad \forall l\in L. \tag{5}$$

The second objective consists of minimizing the number of arcs of the path, $h(p)$, seeking to avoid bandwidth waste, hence favouring global efficiency in the use of network resources as well the reliability of optical connections (longer paths are more prone to failure).

$$\min_{p\in D}\{c_2(p)=h(p)\}. \tag{6}$$

Note that in many cases there is no feasible solution which optimizes the two objective functions, $c_1(p)$ and $c_2(p)$, simultaneously. A certain amount of conflict is therefore expected between c_1 and c_2, and no optimal solution (in most cases) will exist for this problem. Therefore the candidate solutions to the topological RWA multicriteria model are topological paths which are non-dominated solutions to the bi-objective problem:

$$(\mathscr{P})\quad\left\{\begin{array}{l}\min_{p\in D_T} c_1(p)\\ \min_{p\in D_T} c_2(p)\end{array}\right.. \tag{7}$$

Given two paths p_1 and p_2, from s to t in R, path p_1 dominates p_2, denoted by $p_{1D}p_2$, if and only if $c_i(p_1)\leq c_i(p_2)$ $(i=1,2)$ and at least one of the inequalities is strict. A path p is a non dominated solution if no other feasible path dominates it.

The set of admissible solutions, D_T, consists of all topological paths between the source-destination node pair which correspond to *viable lightpaths* p^λ, that is, lightpaths with the same arcs as p and with a free and usable wavelength (according to T_N) in every arc. The topological paths in these conditions (elements of D_T) will be designated as *viable topological paths*, for the given origin-destination node pair. For obtaining D_T firstly the free wavelengths in each arc will have to be identified, taking into account the wavelength conversion capabilities specified in T_N, then the set of viable paths p^λ for each pair of origin-destination nodes becomes implicitly defined.

This model was extended to bidirectional connections between nodes s and t by considering a bidirectional lightpath $p^\lambda=(p_{st}^\lambda,p_{ts}^\lambda)$ supported by a bidirectional topological path $p=(p_{st},p_{ts})$ which is a pair of symmetrical topological paths. In this case the set D_T^b of feasible solutions to the bicriteria model will be the set of viable bidirectional topological paths p, i.e., characterized by the fact that both (unidirectional) topological paths p_{st} and p_{ts} are viable.

Therefore the bi-objective bidirectional routing optimization problem is formulated as:

$$\min_{p\in D_T^b}\left\{c_1(p)=\sum_{l\in p_{st}}\frac{1}{b_l^T}+\sum_{l\in p_{ts}}\frac{1}{b_l^T}\right\} \tag{8}$$

$$\min_{p\in D_T^b}\{c_2(p)=h(p)=h(p_{st})+h(p_{ts})\} \tag{9}$$

We will assume the most common situation in real networks where the two paths p_{st},p_{ts} are topologically symmetrical, thence $h(p)=2h(p_{st})$. Note that this does not imply that the wavelengths used in the two opposite directions are necessarily symmetrical.

2.2. Resolution Approach

The first stage of the resolution approach is an exact algorithm enabling the calculation of non-dominated viable topological paths and the selection of a path according to an automatic procedure that uses preference thresholds defined in the objective function's space and reference points obtained from those thresholds. This algorithmic approach will be reviewed in this subsection.

The second stage is the assignment of wavelengths (and corresponding fibres) along the arcs of the selected path p. For this purpose we will use the maximization of the *wavelength bottleneck bandwidth*, $b_j(p)$ $(\lambda_j\in\Lambda)$:

$$\max_{\lambda_j\in\Lambda}\left\{b_j(p)=\min_{l\in p\wedge b_{lj}>0} b_{lj}\right\}\quad (p\in D_T). \tag{10}$$

Note that this procedure is equivalent to the choice of the least loaded wavelength (LL) along the arcs of the path. Moreover, if all the nodes of the network enable full wavelength conversion, once a viable topological path is selected, the choice of the wavelength(s) in the arcs is irrelevant in terms of network performance. When the nodes have no conversion capability the proposed scheme of wavelength selection is known to give good results (see, e.g., [3]). In any case it can be concluded from many studies that the critical factor in terms of network performance is the selection of topological paths, the choice of wavelength having a minor impact.

In the present model this choice of wavelength will correspond to specify λ_{j^*} in arc l^*:

$$b_{l^*j^*}=\max_{\lambda_j\in\Lambda}\left\{b_j(p)=\min_{l\in p\wedge b_{lj}>0} b_{lj}\right\} : \begin{array}{l}\exists \text{ viable } p^\lambda \text{ which}\\ \text{uses } \lambda_{j^*} \text{ in } l^*\in p\end{array} \tag{11}$$

For bidirectional connections, once a non-dominated solution $p \in D_T^b$ has been selected, the wavelengths (and fibres) to be used along p_{st} and p_{ts} are chosen applying the same procedure to each path. Note that the chosen wavelength(s) in each path can be different.

The aim of the resolution procedure is to find a *good* compromise path from the set of non-dominated solutions, according to certain criteria, previously defined. It must be stressed that path calculation and selection have to be fully automated, as part of a telecommunication network routing mechanism, therefore the use of an interactive decision approach is precluded.

The candidate solutions are computed according to an extremely efficient k-shortest path algorithm, MPS [20], [22], by using a version adapted to paths with a maximum number of arcs (length constrained k-shortest paths) as described in [19]. The algorithm is applied to the convex combination of the two objective functions:

$$f(p) = \alpha c_1(p) + (1-\alpha) c_2(p) \qquad 0 \leq \alpha \leq 1 \,. \tag{12}$$

Note that the value of α just determines the order in which the solutions are found, and its choice is purely instrumental, since all non-dominated solutions can be calculated.

The selection of a solution is based on the definition of preference thresholds for both functions in the form of requested and acceptable values for each of them. These thresholds enable the specification of priority regions in the objective function's space, as illustrated in Fig. 1.

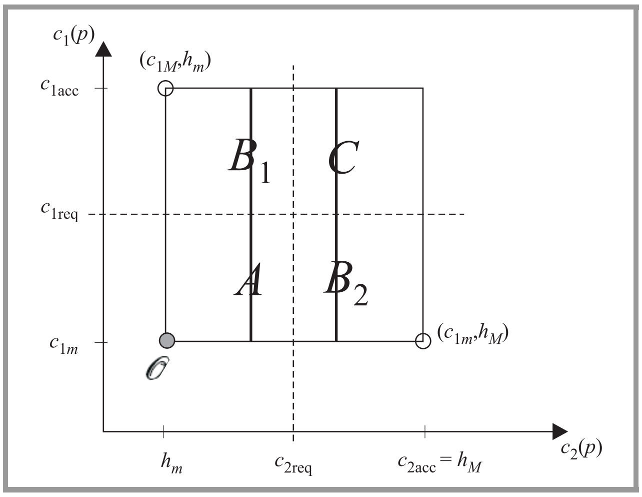


Fig. 1. Preference regions.

In the first step, vertex solutions p^{c_1} and p^{c_2} (viable topological paths) which optimise each objective function, $c_1(p)$ and $c_2(p) = h(p)$, respectively, are computed by solving the associated shortest path problems. This leads to the ideal solution, $\mathcal{O}$, in the objective functions' space.

$$p^{c_1} = \arg\min_{p \in D_T} c_1(p) \tag{13}$$

$$p^{c_2} = \arg\min_{p \in D_T} \{c_2(p) = h(p)\} \tag{14}$$

The preference thresholds $c_{1\mathrm{req}}$, $c_{2\mathrm{req}}$ (requested values) and $c_{1\mathrm{acc}}$, $c_{2\mathrm{acc}}$ (acceptable values) for the two metrics are given, taking into account the discrete nature of $c_2(p) = h(p)$, according to the following expressions:

$$c_{1m} = c_1(p^{c_1}) \quad \wedge \quad c_{2M} = h_M = c_2(p^{c_1}) \,, \tag{15}$$

$$c_{1M} = c_1(p^{c_2}) \quad \wedge \quad c_{2m} = h_m = c_2(p^{c_2}) \,, \tag{16}$$

$$c_{1\mathrm{acc}} = c_{1M} \,, \tag{17}$$

$$c_{2\mathrm{acc}} = h_M \,, \tag{18}$$

$$c_{1\mathrm{req}} = \frac{c_{1m} + c_{1M}}{2} \,, \tag{19}$$

$$c_{2\mathrm{req}} = \left\lfloor \frac{h_m + h_M}{2} \right\rfloor . \tag{20}$$

Priority regions are defined in the objective functions' space according to Fig. 1, in which non-dominated solutions are searched for. Region A (Fig. 1) is the first priority region where the requested values for the two functions are satisfied simultaneously. In the second priority regions, B_1 and B_2, only one of the requested value is satisfied while the acceptable value for the other function is also guaranteed. Concerning these two regions we will give preference to solutions with less arcs, that is, preference is given to solutions in B_1 over solutions in B_2. In region C only the acceptable values, $c_{1\mathrm{acc}}$ and $c_{2\mathrm{acc}}$, are satisfied and it is the last priority region to be searched for.

Finally it will be necessary to select a solution among the non-dominated solutions in the highest priority region, with at least one non-dominated solution, $S \in \{A, B_1, B_2, C\}$. This implies that if no such solutions were found in A, then non-dominated solutions in B_1, B_2, C, in this order would be searched for.

Concerning the selection of a solution when there is more than one non-dominated solution in a given region S, the used method for ordering such solutions is a reference point type approach that considers that the 'form' of the region where solutions are located "reflects" in some manner the user's preferences. At this step a reference point based procedure of the type proposed in [23] is used, by considering as reference point the 'left bottom corner' of region S. This point coincides with the ideal optimum if $S = A$.

Reference type approaches minimize the distance of the solutions to a specific point by using a certain metric, recurring to a scalarizing function [21]. In the present context a weighted Chebyshev metric proportional to the size of the "rectangle" S (see Fig. 2) is used. Therefore, one will select the solution p^*:

$$p^* = \arg\min_{p \in S_N^c} \max_{i=1,2} \{ w_i | c_i(p) - \underline{c}_i | \} \,, \tag{21}$$

where S_N^c is the set of non-dominated paths which correspond to points in S and $(\underline{c}_1, \underline{c}_2)$ is the considered reference point which corresponds to the 'left bottom corner' of region S. The weights w_i of the metrics are chosen in order to obtain a metric with dimension free values:

$$w_i = \frac{1}{\bar{c}_i - \underline{c}_i}, \tag{22}$$

where $(\bar{c}_1, \bar{c}_2)$ is the 'right top corner' of S, so that $\underline{c}_i \leq c_i(p) \leq \bar{c}_i$ $(i = 1, 2)$ for all p such that $(c_1(p), c_2(p)) \in S$. An illustrative example is in Fig. 2, where the number assigned to each bullet is the computation order of the corresponding solution (according to Eq. (12)), and solution (2) would be the one to be selected, since it has the shortest distance to the reference point.

Details of this selection procedure can be seen in [18], [23].

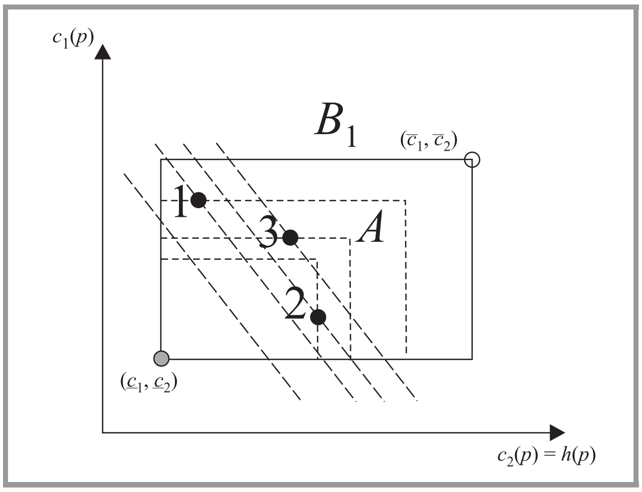


Fig. 2. Choosing the final solution.

In this resolution method, we combine a weighted sum procedure to obtain candidate solutions with a reference-point based method to select a solution in a higher priority region. In this form we sought to make the most of the very great efficiency of the used shortest path ranking procedure, based on the MPS algorithm [22] and of the inherent superiority of the use of a reference point-based procedure, as a solution selection method. Furthermore, note that in the present context, the computational efficiency is a very important aspect taking into account the automated nature of the routing mechanism, which requires a solution in a very short time period. This factor becomes more critical in networks of higher dimension.

The final step of the resolution method is the choice of wavelengths along the arcs of the selected path. This steps follows the procedure described in Subsection 2.2 which is based on the maximization of the wavelength bottleneck bandwidth.

The described resolution method can be applied straightforwardly to the calculation and selection of bidirectional lightpaths, considering the necessary adaptations to the objective functions, specified in Eqs. (8) and (9).

3. Performance Analysis of the Model

Extensive simulations with the model were made on several typical WDM networks found in literature. This section presents the simulation results for five such networks, namely, the NSFNET [24] (see Fig. 3), the Pan-European network COST 266BT [24], [25] (Fig. 4), the denser version of this network [25] – COST 266TT network (see Fig. 5), a typical core network presented in [26] – KL network (Fig. 6), and a typical network provider network presented in [27] – ISP network (Fig. 7). Table 1 summarizes the main characteristics of these networks. All the networks were dimensioned for about one thousand bidirectional lightpaths (1084 for NSFNET, 1008 for both COST 266BT and Cost 266TT, 1050 for KL network, and 918 for ISP network) and each fibre has 16 wavelengths.

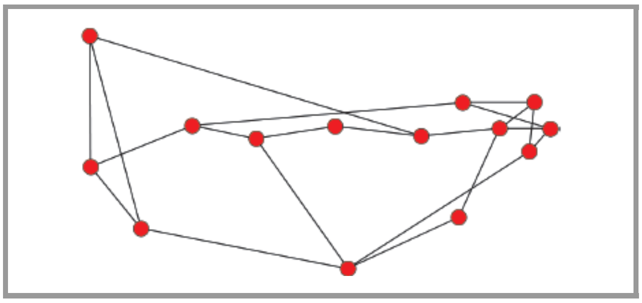


Fig. 3. NSFNET network (14 nodes and 21 links) [24].

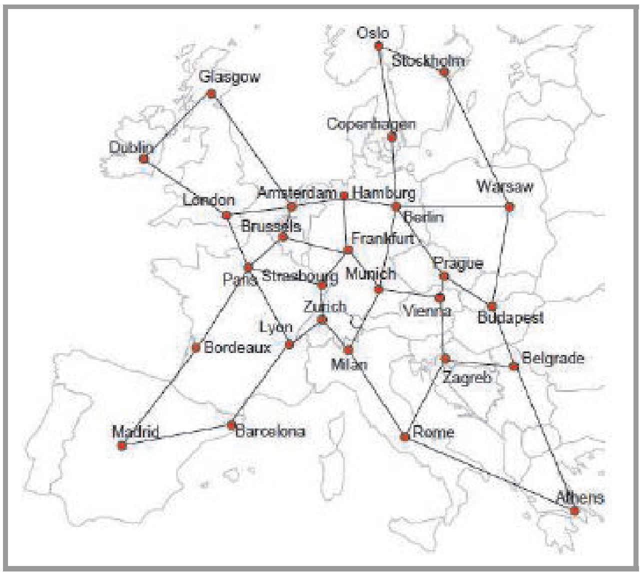


Fig. 4. COST 266BT Pan-European network (28 nodes and 41 links) [24], [25].

Concerning the wavelength conversion aspect, simulations were conducted considering two different scenarios: all nodes without conversion capability and five nodes with total conversion capability (central nodes were chosen with this capability).

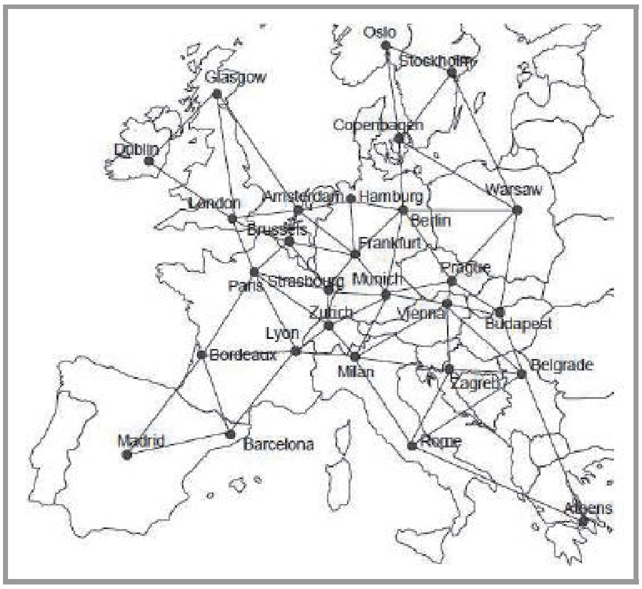


Fig. 5. COST 266TT network (28 nodes and 61 links) [25].

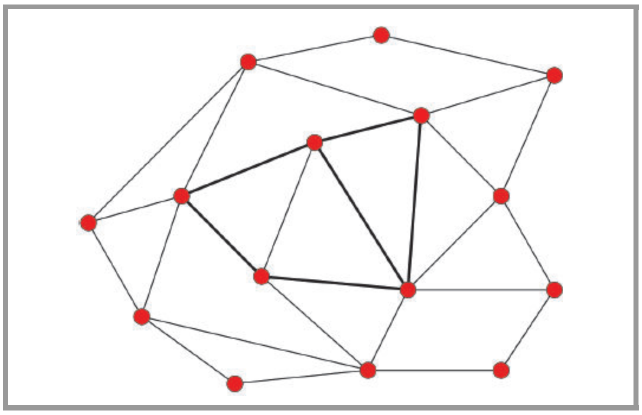


Fig. 6. KL network (15 nodes and 28 links) [26].

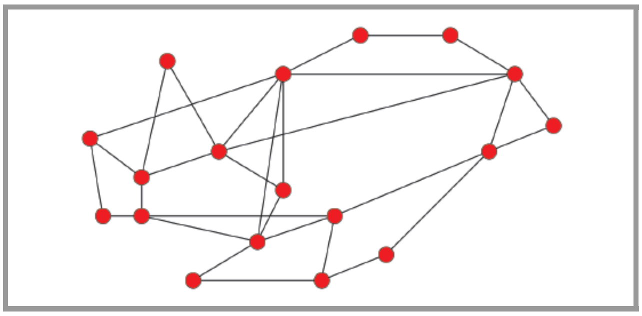


Fig. 7. ISP network (18 nodes and 30 links) [27].

Table 1
Networks characteristics

Network	Number of		Nodal degree
	nodes	links	
NSFNET [24]	14	21	3.00
COST266BT [24, 25]	28	41	2.93
COST266TT [25]	28	61	4.36
KL [26]	15	28	3.73
ISP [27]	18	30	3.33

Simulations considered 1200 connection requests (incremental traffic) in two different cases: with 100% bidirectional requests and with 5% unidirectional requests (because most of the connection requests for lightpaths are bidirectional).

Simulation results showed that the performance of the networks where five nodes have total conversion capability is nearly the same as for the networks without conversion. Therefore, from now on, we only present the "no conversion" scenario. Discussion and conclusions remain true for the second scenario.

For performance assessment purposes, results obtained using the bicriteria (BiC) model will be compared with the corresponding results using the single objective formulations, namely, the first objective function related with the bandwidth usage (SP_c1), and the shortest path concerning hop count (SP_c2). Several relevant network performance measures will be used in this comparison.

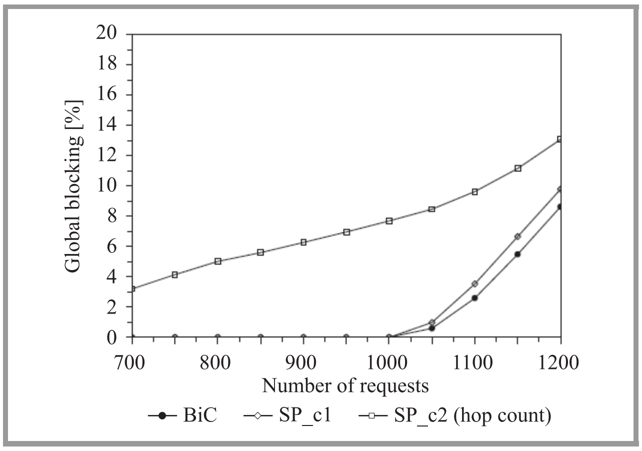


Fig. 8. Global blocking – NSFNET network.

Figure 8 shows the global blocking probability in the NSFNET network, for 100% bidirectional requests. As we can see, the bicriteria approach leads to a blocking probability lower than that in the single objective formulations. The difference is significantly higher with the shortest path approach SP_c2. Also note that until 1000 requests both BiC and SP_c1 do not exhibit any blocking. The SP_c2 model rejects requests much earlier, and this clearly confirms that choosing the shortest path based only on the hop count is a poor strategy.

In Fig. 9, the number of accepted requests is shown together with the used bandwidth (BW). When the number of requests is high, the used bandwidth exceeds 95%, but this corresponds to approximately 1100 accepted requests in BiC, a value that surpasses the number of connections for which the network was dimensioned (1084 for the NSFNET).

Although the BiC model uses more bandwidth than the SP_c2, it should be noted that BiC supports a significantly higher number of connections. In fact, as it can be seen in Fig. 10, BiC allows a lower average number of hops per connection. Another interesting conclusion that emerges from the analysis of Fig. 9 is that the BiC, while

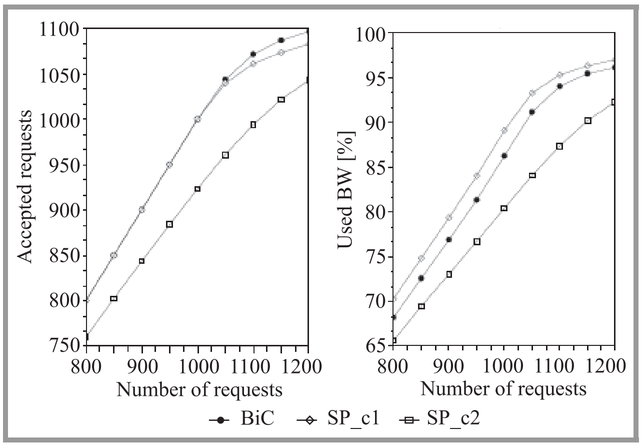


Fig. 9. Accepted requests versus used bandwidth – NSFNET network.

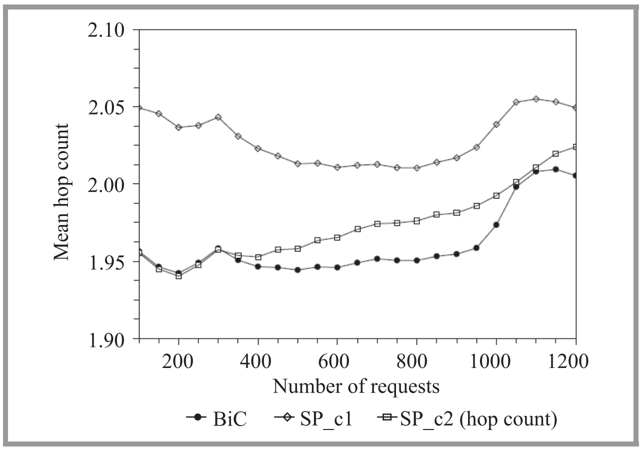


Fig. 10. Mean hop count – NSFNET network.

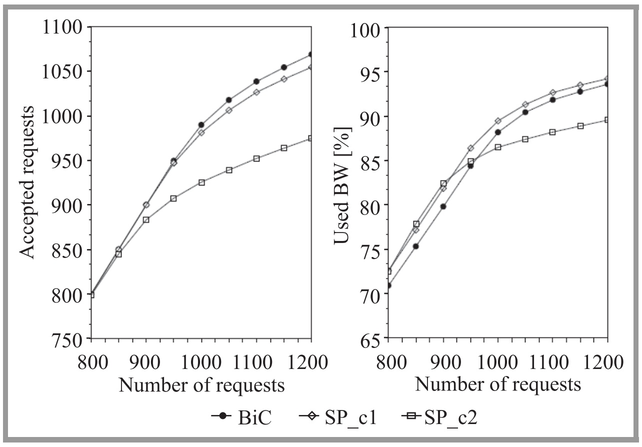


Fig. 11. Accepted requests versus used bandwidth – COST 266BT network.

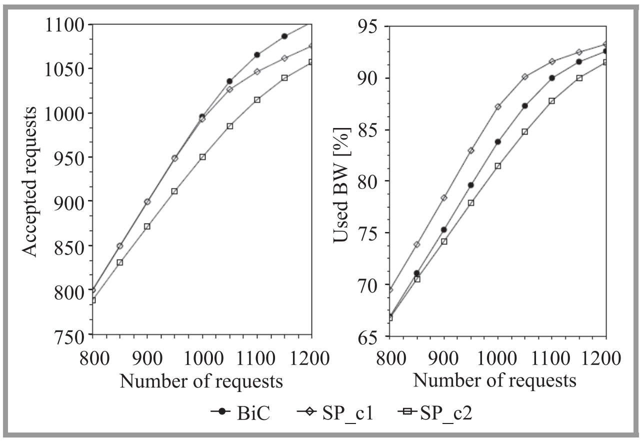


Fig. 12. Accepted requests versus used bandwidth – COST 266TT network.

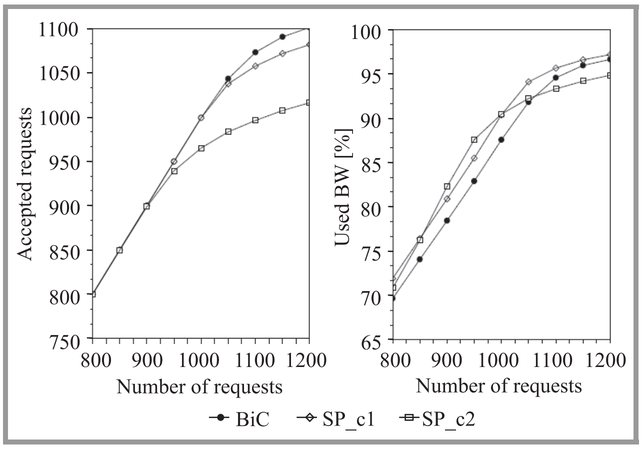


Fig. 13. Accepted requests versus used bandwidth – KL network.

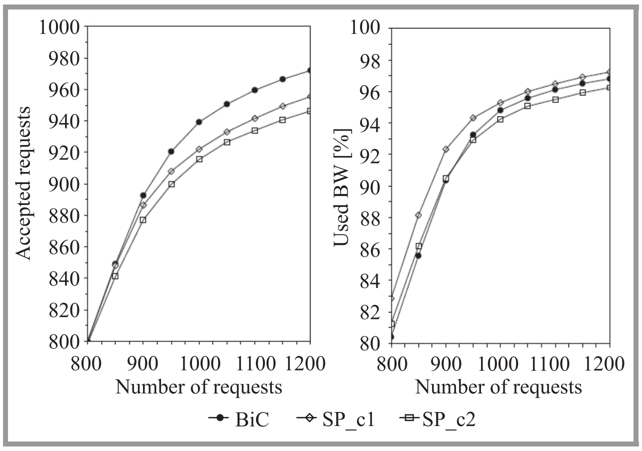


Fig. 14. Accepted requests versus used bandwidth – ISP network.

accepting more requests than SP_c1, when traffic load is high, always uses less bandwidth, which shows its superior performance.

Although not shown here, the results obtained when 5% of the requests were unidirectional are rather similar to the ones with 100% bidirectional connections.

The results in the other networks exhibit the same behavior. In the COST 266BT network, BiC also has lower blocking than SP_c1 and SP_c2 (more accepted requests) but always uses an amount of bandwidth smaller than SP_c1 (see Fig. 11). For moderated traffic loads (until 950 connection requests), BiC even uses less bandwidth

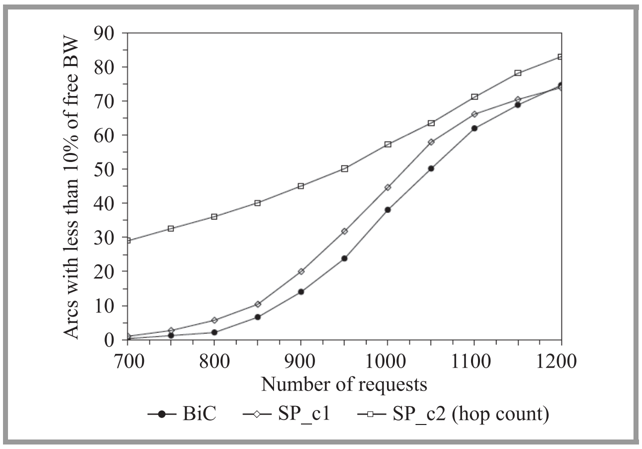


Fig. 15. Arcs with less than 10% of free bandwidth – COST 266TT network

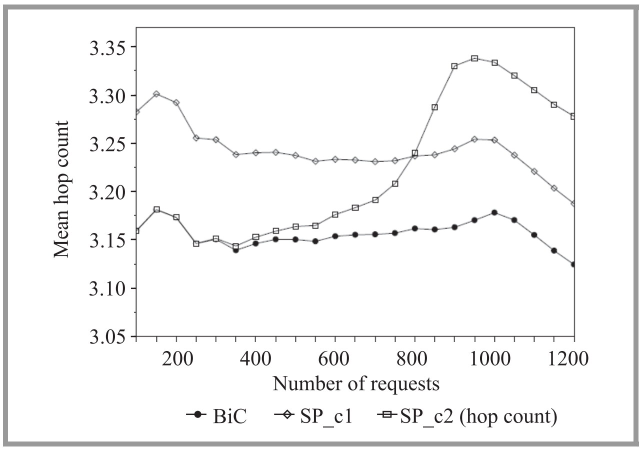


Fig. 16. Mean hop count – COST 266BT network.

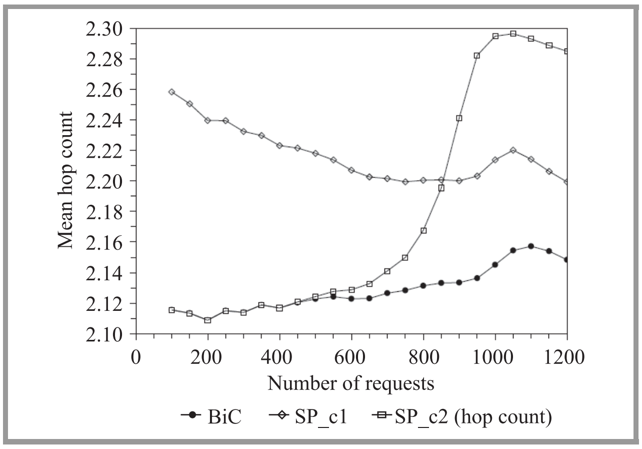


Fig. 17. Mean hop count – KL network.

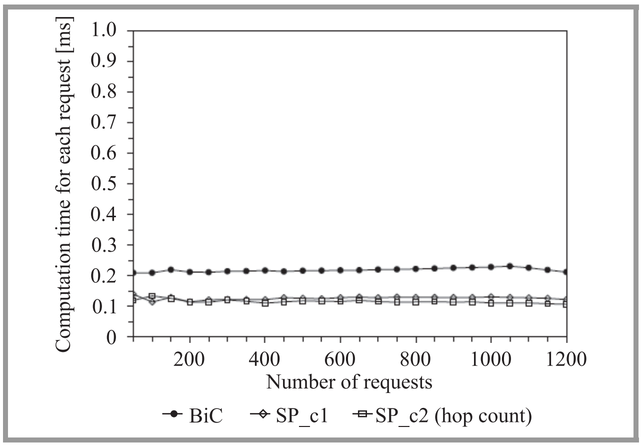


Fig. 18. Computation time for each request – NSFNET network.

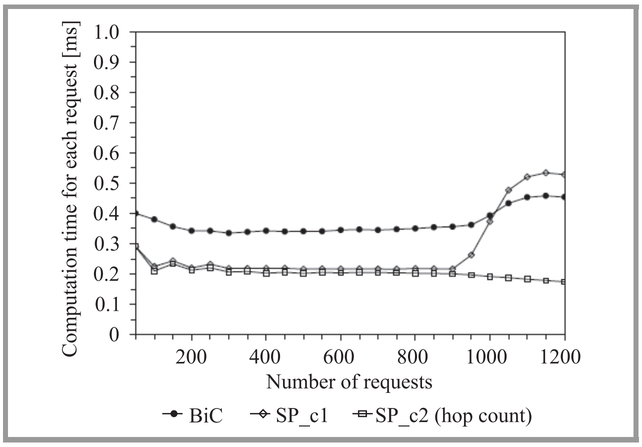


Fig. 19. Computation time for each request – COST 266BT network.

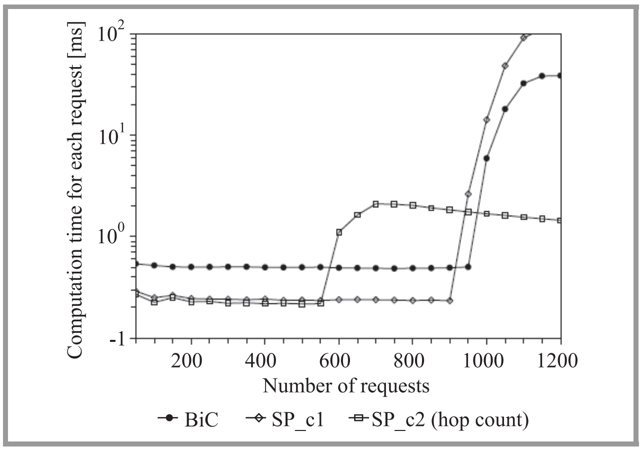


Fig. 20. Computation time for each request – COST 266TT network.

than SP_c2, despite allowing the establishment of many more lightpaths.

In Figs. 12, 13, and 14 the same performance measures are shown for COST 266TT, KL, and ISP networks, respectively. The superior performance of the BiC model is consistent in all tested networks.

It is also interesting to analyze the ability of the different approaches to distribute traffic over the network. Figure 15 plots the number of links in the COST 266TT network with less than 10% of free bandwidth. BiC leads to a lower number of links with less than 10% free bandwidth than SP_c1 and SP_c2. Knowing that the number of accepted request is also higher, we can conclude that BiC has a performance significantly better than the single objective counterparts. The same behavior was observed in the remaining simulated networks (although not shown here).

Another interesting network performance measure is the average number of hops per established lightpath. As it can be seen in Figs. 10, 16, and 17 (for NSFNET, COST 266BT and KL networks, respectively), BiC uses in average a smaller number of links in all ranges of traffic loads. This happens because it takes advantage of the characteristics of the two metrics. For light traffic, the BiC model chooses shorter connections and, in fact, achieves paths as short as SP_c2. But, unlike SP_c2, BiC is concerned with the load already present in the network links. On the other hand, SP_c1 does not take into account the hop count, leading to longer paths, even when the network is nearly "empty". As the traffic load increases, the worst choice of the initial paths in SP_c2 leads to bottlenecks in some links. This results in the selection of longer paths, and in higher blocking probability. Above approximately 800 requests, the average number of hops per lightpath in SP_c2 is even greater than in SP_c1 – the traffic distribution is more effective in this model. Also note that when the number of connection requests exceeds approximately 1000, the mean hop count decreases in the three approaches. With this traffic load the network is already congested and node pairs topologically distant are experiencing greater difficulties in establishing a successful connection. So only some "short" connection requests obtain a service, lowering the mean hop count.

Regarding CPU time, the BiC approach requires more CPU, as would be expected, but CPU times are still very low, not exceeding 0.25 ms in NSFNET, 0.45 ms in COST 266BT, 0.3 ms in KL network, and 0.35 ms in ISP network. In the COST 266TT network CPU time remains under 0.6 ms until 950 connection requests. Figures 18, 19, and 20 show the CPU time per connection request for NSFNET, COST 266BT and COST 266TT networks, respectively. The CPU times remain stable as the traffic load grows in the NSFNET, ISP, KL and COST 266BT networks. In COST 266TT network (see Fig. 20), above 950 connection requests CPU time for BiC and SP_c1 approaches increases considerably, and can be as high as 40 ms. This effect occurs when the traffic load is very high and coincides with the starting of visible blocking in the network. Note that this substantial increase in CPU time is even larger in SP_c1.

In order to assess the degree of conflict between the two objective functions used in BiC approach, the number of requests without an optimal solution was calculated. Figures 21 and 22 show the percentage of non-dominated non-optimal solutions in COST 266BT and COST 266TT networks. Although the number of non-dominated solutions is relatively low this does not compromise the interest in using a bicriteria model. In fact many of the ideal solutions of the bicriteria model might possibly have not been found by the single objective models because they correspond to alternative optimal solutions in one of the objective functions.

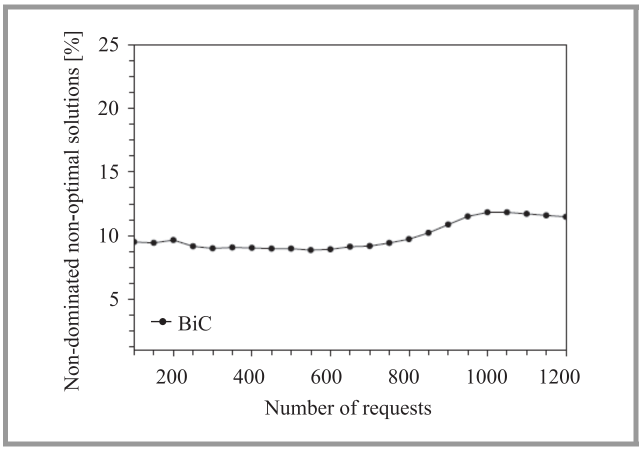


Fig. 21. Non-dominated non-optimal solutions – COST 266BT network.

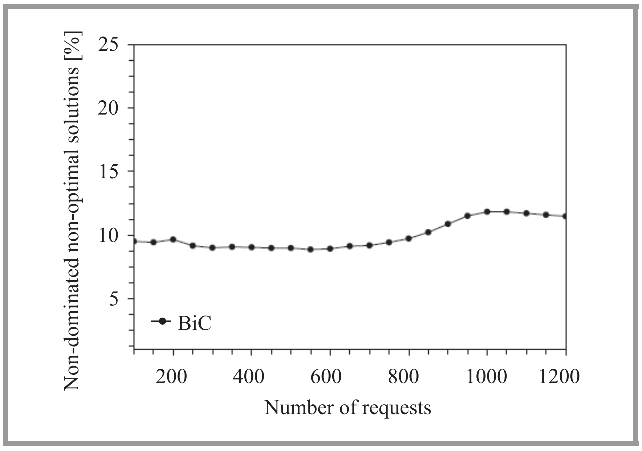


Fig. 22. Non-dominated non-optimal solutions – COST 266TT network.

4. Conclusions

The routing and wavelength assignment problem in WDM networks, as seen from a full traffic engineering perspective, involves multiple metrics, to be optimized, and specific constraints. Therefore multicriteria approaches like the one described in this paper enable to explicitly represent the various performance objectives and to address, in a consistent manner, the trade offs among the various criteria.

A bicriteria model for obtaining a topological path (unidirectional or symmetric bidirectional) for each lightpath request in a WDM network was reviewed. The model considers two criteria – the first one takes into account the bandwidth usage in the links of the network and the second one the number of links of the path. The automated resolution approach uses a *k*-shortest path algorithm, as well as preference thresholds defined in the objective function's space, combined with a Chebyshev distance to a reference point (which changes with the analyzed preference region). Having obtained a non-dominated topological path, a heuristic procedure was then used to assign wavelengths to the links.

The performance of this bicriteria model was analyzed using several benchmark networks, and considering a comparison with the results of the two single criterion approaches corresponding to each of the criteria used in the BiC model. Clearly, the BiC approach resulted in lower global blocking than the single criterion models SP_c1 and SP_c2. This is due to an initial better choices of paths and a more balanced distribution of traffic load. At moderate load, although BiC approach accepts more requests, BiC uses less bandwidth than SP_c1; SP_c2 uses less bandwidth than the BiC but it leads to a significant lower number of successful connections.

The impact of having five nodes with wavelength conversion capability was negligible in the simulated situations.

Although the BiC approach uses more CPU time per request its performance was nevertheless quite good – below 0.5 ms except in the denser network (COST 266TT).

In a following paper we will address the network performance analysis issues of an extension of the bicriteria model that provides dedicated path protection, in the event of failures. This is an issue of great importance having in mind the great amount of traffic carried in these optical networks. To provide the necessary network resiliency, while preserving the multicriteria nature of the developed model, this extension (see [28]) enables to obtain a topological pair of node disjoint paths for each connection request. The developed performance analysis study will consider the same type of experimentation and performance measures as in this paper and will present interesting conclusions concerning the potentialities (put in evidence in this paper) and limitations of the use of multicriteria approaches in this context.

Acknowledgements

This study was partially supported by PT Inovação and by programme POSC of the EC programme cosponsored by national funds.

Carlos Simões also thanks the FCT and IPV by the grant of the Ph.D. scholarship SFRH/BD/49739/2009.

References

[1] X. Chu, J. Liu, and Z. Zhang, "Analysis of sparse-partial wavelength conversion in wavelength-routed WDM networks", in *Proc. IEEE Conf. INFOCOM'2004*, Hong Kong, China, 2004, vol. 2, pp. 1363–1371.

[2] S. Bandyopadhyay, A. Jaekel, A. Sengupta, and W. Lang, "A virtual wavelength translation scheme for routing in all-optical networks", *Phot. Netw. Commun.*, vol. 4, pp. 391–407, 2002.

[3] H. Zang, J. P. Jue, and B. Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks", *Opt. Netw. Mag.*, vol. 1, pp. 47–60, 2000.

[4] R. Ramaswami and K. Sivarajan, "Routing and wavelength assignment in all-optical networks", *IEEE/ACM Trans. Netw.*, vol. 3, pp. 489–500, 1995.

[5] I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: An approach to high bandwidth optical WAN's", *IEEE Trans. Commun.*, vol. 40, pp. 1171–1182, 1992.

[6] D. Banerjee and B. Mukherjee, "A practical approach for routing and wavelength assignment in large wavelength-routed optical networks", *IEEE J. Sel. Area Commun.*, vol. 14, pp. 903–908, 1996.

[7] A. E. Ozdaglar and D. P. Bertsekas, "Routing and wavelength assignment in optical networks", *IEEE/ACM Trans. Netw.*, vol. 11, pp. 259–272, 2003.

[8] B. Jaumard, C. Meyer, and B. Thiongane, "Comparison of ILP formulations for the RWA problem", *Opt. Switch. Netw.*, vol. 4, pp. 157–172, 2007.

[9] R. M. Krishnaswamy and K. N. Sivarajan, "Algorithms for routing and wavelength assignment based on solutions of LP-relaxations", *IEEE Commun. Lett.*, vol. 5, pp. 435–437, 2001.

[10] M. Saad and Z.-Q. Luo, "On the routing and wavelength assignment in multifiber WDM networks", *IEEE J. Sel. Area Commun.*, vol. 22, pp. 1708–1717, 2004.

[11] R. Krishnaswamy and K. Sivarajan, "Design of logical topologies: A linear formulation for wavelength-routed optical networks with no wavelength changers", *IEEE/ACM Trans. Netw.*, vol. 9, pp. 186–198, 2001.

[12] B. Jaumard, C. Meyer, and B. Thiongane, "ILP formulations for the routing and wavelength assignment problem: symmetric systems", in *Handbook of Optimization in Telecommunications*. Springer Science + Business Media, 2006, pp. 637–677.

[13] J. Clímaco and J. Craveirinha, "Multicriteria analysis in telecommunication planning and design – problems and issues", in *Multiple Criteria Decision Analysis: State of the Art Surveys*, vol. 78 of *International Series in Operations Research & Management Science*. New York: Springer Science, 2005, pp. 899–951.

[14] J. Clímaco, J. Craveirinha, and M. Pascoal, "Multicriteria routing models in telecommunication networks – overview and a case study", in *Advances in Multiple Criteria Decision Making and Human Systems Management*. IOS Press, 2007, pp. 17–46.

[15] A. Todimala and B. Ramamurthy, "Congestion-based algorithms for online routing in optical WDM mesh networks", in *Proc. IASTED Int. Conf. CIIT'03*, Scottsdale, USA, 2003, pp. 43–48.

[16] X. Chu and B. Li, "Dynamic routing and wavelength assignment in the presence of wavelength conversion for all-optical networks", *IEEE/ACM Trans. Netw.*, vol. 13, pp. 704–715, 2005.

[17] C.-F. Hsu, T.-L. Liu, and N.-F. Huang, "On adaptive routing in wavelength-routed networks", *Opt. Netw. Mag.*, vol. 3, pp. 15–24, 2002.

[18] T. Gomes, J. Craveirinha, J. Clímaco, and C. Simões, "A bicriteria routing model for multi-fibre WDM networks", *Phot. Netw. Commun.*, vol. 18, pp. 287–299, 2009.

[19] T. Gomes, L. Martins, and J. Craveirinha, "An algorithm for calculating the *k* shortest paths with a maximum number of arcs", *Investigação Operacional*, vol. 21, no. 2, pp. 235–244, 2001.

[20] E. Martins, M. Pascoal, and J. Santos, "Deviation algorithms for ranking shortest paths", *Int. J. Found. Comput. Sci.*, vol. 10, no. 3, pp. 247–263, 1999.

[21] A. P. Wierzbicki, "The use of reference objectives in multiobjective optimization", *Theory and Application, Proceedings, Lecture Notes in Economics and Mathematical Systems*, vol. 177, pp. 468–487, 1980.

[22] E. Martins, M. Pascoal, and J. Santos, "An algorithm for ranking loopless paths", Tech. Rep. 99/007, CISUC, 1999 [Online]. Available: http://www.mat.uc.pt/~marta/Publicacoes/mps2.ps

[23] J. Clímaco, J. Craveirinha, and M. Pascoal, "An automated reference point-like approach for multicriteria shortest path problems", *J. Syst. Sci. Syst. Engin.*, vol. 15, pp. 314–329, 2006.

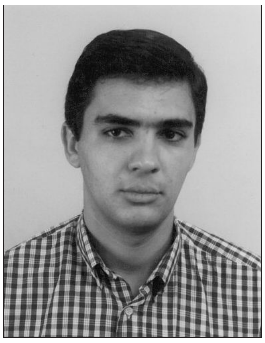
[24] A. Betker, C. Gerlach, M. Jäger, M. Barry, S. Bodamer, J. Späth, C. M. Gauger, and M. Köhn, "Reference transport network scenarios MultiTeraNet Report", 2003, [Online]. Available: http://iospress.metapress.com

[25] S. Maesschalck, D. Colle, M. Lievens, P. Demeester, C. Mauz, M. Jaeger, R. Inkret, B. Mikac, and J. Derkacz, "Pan-European optical transport networks: an availability-based comparison", *Phot. Netw. Commun.*, vol. 5, pp. 203–225, 2003.

[26] M. Kodialam and T. Lakshman, "Dynamic routing of bandwidth guaranteed tunnels with restoration", in *Proc. INFOCOM'2000 Conf.*, Tel-Aviv, Israel, 2000, vol. 2, pp. 902–911.

[27] G. Apostolopoulos, R. Guérin, S. Kamat, and S. Tripathi, "Quality of service based routing: a performance perspective", in *Proc. ACM SIGCOMM '98 Conf,*, Vancouver, Canada, 1998, pp. 17–28.

[28] T. Gomes, C. Simões, J. Craveirinha, and J. Clímaco, "A bi-objective model for routing and wavelength assignment in resilient WDM networks", in *Safety, Reliability and Risk Analysis: Theory, Methods and Applications*, vol. 4, pp. 2627–2634, 2008.



Carlos Simões obtained an undergraduate diploma in electrical engineering science- telecommunications and electronics in 1995 and a M.Sc. degree in systems and automation-telecommunications in 1999, both from the University of Coimbra, Portugal. Since 1998, he has been an Assistant at the Polytechnic Institute of Viseu, Portugal, and a researcher at INESC-Coimbra R&D Institute. He is currently working towards his Ph.D. degree in electrical engineering and informatics at the University of Coimbra and is involved in the research of multicriteria models for routing in optical networks.
e-mail: csimoes@ipv.pt
School of Technology and Management of Viseu
Polytechnic Institute of Viseu
Campus Politecnico de Repeses
P-3504-510 Viseu, Portugal

Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal

João Clímaco is Full Professor at the Faculty of Economics of the University of Coimbra and President of the Scientific Committee of the INESC-Coimbra. He obtained the M.Sc. degree in control systems at the Imperial College of Science and Technology, University of London (1978); the "Diploma of Membership of the Imperial College of Science and Technology" (1978); the Ph.D. in optimization and systems theory, Electrical Engineering Department, University of Coimbra (1982) and the title of "Agregação" at the University of Coimbra (1989). He served in the past as the Vice-President of ALIO- Latin Ibero American OR Association and Vice-President of the Portuguese OR Society. He belongs to the editorial board of the following scientific journals: Investigação Operacional (Journal of the Portuguese OR Society), Journal of Group Decision and Negotiation, International Transactions in Operational Research (ITOR), ENGEVISTA and Rio's International Journal on Sciences of Industrial and Systems Engineering and Management. He is also member of the editorial board of the University of Coimbra Press. His current interests of research include multicriteria decision aiding, multiobjective mathematical programming, location analysis and telecommunication network planning and management.
e-mail: jclimaco@inescc.pt
Faculty of Economics University of Coimbra
Av. Dias da Silva 165
P-3004-512 Coimbra, Portugal

Faculty of Economics
Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal

Teresa Gomes, José Craveirinha – for biographies, see this issue, p. 12.

Paper

Performance Analysis of a Bi-Objective Model for Routing with Protection in WDM Networks

Carlos Simões, Teresa Gomes, José Craveirinha, and João Clímaco

Abstract—The operation of wavelength division multiplexing (WDM) networks involves not only the establishment of lightpaths, defining the sequence of optical fibres and the wavelength in each fibre for traffic flow, but also a fault management scheme in order to avoid the huge loss of data that can result from a single link failure. Dedicated path protection, which establishes two end-to-end disjoint routes between the source–destination node pair, is an effective scheme to preserve customers' connections. This paper reviews a bicriteria model for dedicated path protection, that obtains a topological path pair of node-disjoint routes for each lightpath request in a WDM network, developed by the authors. An extensive performance analysis of the bicriteria model is then presented, comparing the performance metrics obtained with the monocriterion models using the same objective functions, in four different reference networks commonly used in literature.

Keywords—multicriteria optimization, protection, routing in WDM networks.

1. Introduction

1.1. Background Concepts

In modern all-optical networks based on wavelength division multiplexing (WDM), one single fiber can provide an enormous bandwidth (up to tens of terabits per second) by multiplexing many non-overlapping wavelength channels. Each wavelength can be operated transparently, at speeds compatible with the lower capacity of the end-users devices.

The high capacity of a single fibre in optical networks, however, has the drawback that a failure on a link can potentially lead to a huge amount of data loss (and revenue), and service disruption for a large number of customers. In this scenario, network survivability becomes a critical concern for service providers (both in the network design phase and in the real-time network operation) and fast and efficient fault-recovery mechanisms are then needed to ensure a high degree of network resilience and minimize losses. Survivability of a network refers to the network capability to provide continuous service in the presence of failures.

Fibre cuts are usually the most frequent failure event in optical networks, and lead to the disruption of all the lightpaths that transverse the failed fibre. But other network equipments (such as OXC, amplifiers, etc.) may also fail. These two basic types of failures in the network can be categorized as either link (mostly cable cuts) or node failures (equipment malfunctions).

Essentially, there are two types of fault-recovery mechanisms. A lightpath can be protected against failures by pre-computing a backup route and reserving resources along the route in advance [1]. We call this approach a protection scheme. Alternatively, the resources necessary to restore a disrupted lightpath can be discovered dynamically and signaled (reserved) only after a failure occurs. This approach is referred to as dynamic restoration (or just restoration) [1]. Usually, dynamic restoration schemes are more resource-efficient because they do not allocate spare capacity in advance and provide resilience against different kinds of failures (including multiple failures), but they need more time to discover free resources and reroute the disrupted connection. On the other hand, a protection scheme has faster recovery time and can guarantee resource availability for a backup path in the fault scenarios for which it was designed [2], but it needs more resources.

A protection method can protect the end-to-end path (path protection), protect the failed link (link protection) or protect a segment of a path (subpath protection) [2]. In path protection, in order to recover from any single link failure in the network, a link-disjoint path is needed as the backup path to reroute the traffic on the active path (primary path). The primary and backup paths for a connection between a node pair must be link disjoint so that no single link failure can affect both paths. Note that node failures can also be considered by calculating node disjoint routes. In link protection, the traffic is rerouted only around the failed link. While path protection leads to efficient utilization of backup resources and lower end-to-end propagation delay for the recovered route, link protection provides shorter protection switching time. The concept of subpath protection has been proposed as a tradeoff between the path and link protection schemes, and consists in the division of the primary path into a sequence of segments, each one protected separately [3]. In dedicated protection there is no sharing between backup resources, while in shared protection backup wavelengths can be shared on some links as long as their protected segments (links, subpaths, paths) are mutually diverse. Consequently, shared protection is more resource efficient, but the backup paths can not be configured until the failure occurs and, thus, recovery time is longer than with dedicated protection.

1.2. Routing and Wavelength Assignment

A lightpath may span several fibre links and consist of wavelength channels in the sequence of these links, interconnected at the nodes by means of optical routing. In order to establish a lightpath, the network needs to decide on the topological route and the wavelength(s) for the lightpath. If the optical cross-connects have wavelength converters (wavelength-convertible network), a lightpath can be assigned to different wavelengths in each link of its route. However, since wavelength converters are costly and may cause signal quality degradation, often no wavelength converters are used or only some nodes have this capability. In the absence of wavelength conversion (wavelength-continuous network), the same wavelength must be allocated on all links in the path (*the wavelength continuity constraint*), but wavelengths can be reused by different lightpaths in the network, as long as they do not share any fibre.

Given a set of connection requests, the routing and wavelength assignment (RWA) problem consists of deciding the path and assign a wavelength to each of its links, for every request, given a desired objective and a set of constraints [4]. Wavelength assignment must satisfy two constraints, namely, no two lightpaths on the same physical link can be assigned the same wavelength, and if wavelength conversion is not available, then wavelength continuity constraint must be satisfied on all the links that a lightpath traverses.

Obviously, wavelength conversion leads to lower blocking probabilities, but, in practice, some works have shown that with only a small number of converters placed in strategic locations, a significant performance improvement can be achieved [5].

The RWA problem is known to be NP-complete [6]. Hence, most approaches presented in the literature decouple the problem into its two underlying sub-problems – routing and wavelength assignment – which are solved separately. However each sub-problem is still NP-complete [6]. Therefore, the proposed methods in the literature are generally based on heuristics that allow obtaining a feasible solution in acceptable computation time. Generally, the routing scheme has a much a higher impact in the blocking probability of the connections than the wavelength assignment scheme [4].

1.3. Survivable Routing and Wavelength Assignment

In a WDM network employing path protection, the problem of finding a disjoint primary-backup path pair and assigning wavelength(s) to each path is known as the survivable routing and wavelength assignment (S-RWA) problem and has been extensively studied [1], [2], [7], [8], [9].

Typically, routing heuristics prefer the path pair with least cost from a source to a destination to carry the traffic, where the path cost is defined to be the sum of the costs of all the links along the path. The path cost of a dedicated path-protected connection is the sum of the costs of the primary and backup lightpaths.

Concerning shared path protection, the path cost of a connection is the sum of the cost of the primary lightpath and the costs of the additional backup links on which the wavelength is reserved but is not shared by existing connections. The path pair can be either selected from a set of preplanned alternate routes or dynamically computed according to current network state. Depending on different traffic engineering considerations, different cost functions can be applied to network links, such as constant 1 (to minimize hop distance), length of the links (to minimize delay), fraction of available capacity on the links (to balance traffic load), network cost (total equipment cost plus operational cost) on the links (to minimize cost), and so on. Wavelength assignment can be considered only after the routing of the primary-backup path pair. Several wavelength assignment heuristics have been proposed in the literature [4]. Wavelength assignment can also be jointly considered with the route computation of both primary and backup paths.

In dedicated path protection, two disjoint routes are needed between the source node and the destination node – one for the primary path and the other for the backup path. The simplest way to compute disjoint paths consists in two steps [7]–[10]. In the first step the primary path is computed using a shortest-path algorithm. Then, in the second step, the links and nodes used in the primary path are removed and the backup path is calculated in the remaining topology. This approach is referred to as the two-step approach and has some drawbacks because of the sequential nature of paths' calculation. First, although the primary path is the shortest one (minimal cost), the sum of the costs of the two disjoint paths may not be optimal (minimal). Worst than that, in some scenarios, since erasing the first path can isolate the source node from the destination node, this procedure may not find a pair of disjoint paths even if such a pair of paths exist (trap topology). This can happen even in highly connected topologies [10].

To find two disjoint paths with minimal total cost, Suurballe's algorithm [11] can be applied. This algorithm guarantees to find the disjoint path-pair in polynomial time if such pair exists.

1.4. Multicriteria Models

Typically, routing protocols try to optimize a single metric, using some variant of a shortest path algorithm. Nevertheless, all-optical WDM networks can be characterized in terms of performance by multiple metrics, and the design of real networks usually involves multiple, often conflicting objectives and various constraints. Clearly, since single objective approaches can not express this multiplicity of metrics, it seems potentially advantageous to develop multicriteria models that explicitly represent the different performance objectives, enabling to treat in a consistent manner the trade off among the various criteria.

Note that in models involving explicitly multiple criteria, there is no guarantee that a solution that optimizes all the criteria exists, and the concept of optimal solution is replaced by the concept of non-dominated solutions. A non-dominated solution is a feasible solution such that no improvement in any criterion may be achieved without sacrificing at least one of the other criteria.

Reference [12] presents a state-of-art review on multicriteria approaches in communication networks, including a section dedicated to routing models. For a more recent review on multicriteria routing models see [13].

A bicriteria model for obtaining a topological path (unidirectional or symmetric bidirectional) for each lightpath request in a WDM network with multi-fibre links and an exact resolution approach for that model was presented by the authors in [14], and an extensive performance analysis of the bicriteria model in several reference WDM networks can be found in [15]. In order to provide dedicated path protection to lightpaths, against node failures, an extension of the bicriteria model that allows to obtain a topological pair of node disjoint paths for each request was developed in [16]. The first criterion is related to bandwidth usage in the links of the network, and the second criterion is the number of links (hops) of the path. The resolution approach of this model uses a *k*-shortest path algorithm for the determination of non-dominated shortest pairs of disjoint paths proposed in [17]. Furthermore, preference thresholds, defined in the objective function's space, combined with a Chebyshev distance to a reference point [18] are used for selecting the final solution. The solution of this bicriteria model is a non-dominated topological (optically feasible) disjoint path pair. A heuristic procedure is then used to assign the wavelengths in the links of the two disjoint paths.

In this paper we focus on the problem of dedicated path protection against node failures, and present an extensive and systematic performance analysis study of the bicriteria model with dedicated protection developed in [16]. This analysis considers relevant network performance measures and compares the corresponding results for the bicriteria model with the results of the associated single objective models, one related to the bandwidth usage and the other consisting of the total number of links in the two paths (active and protection path). An incremental traffic model (where the duration of the connections is assumed unlimited) and several benchmark networks commonly used in this research area will be considered. Essentially, the network performance measures envisaged are: the frequency of rejected requests (an estimate of the global blocking probability), the total used bandwidth, the mean hop count of accepted requests, the percentage of links with minimal free bandwidth, the average CPU time per request, and the percentage of non-optimal solutions.

The paper is organized as follows. In Section 2 the model with dedicated protection is described, together with the resolution approach of the bicriteria model. Performance analysis of the results obtained using several network topologies are presented and discussed in Section 3, enabling to compare the network performance (under the prescribed metrics) of the bicriteria with the monocriterion models, with dedicated protection. Finally, some conclusions of practical and methodological nature are drawn in Section 4.

2. The Bicriteria Routing Model with Dedicated Protection

2.1. Model Description

In this section we describe the features of the proposed bicriteria routing model associated with the dynamic lightpath establishment problem (DLE) with incremental traffic, and a mixture of unidirectional and bidirectional (symmetric) connections requests, in WDM networks. The model was developed for application in large WDM networks, with multiple wavelengths per fibre and multi-fibres per link. In order to cover a wide variety of networks, different types of nodes are considered (with complete wavelength conversion capability, limited range conversion or no wavelength conversion capability) in the model. Due to the real-time nature of the intended application, solutions should be obtained in a short time. This requirement lead to the separation of the routing and wavelength assignment problems, having in mind an automatic selection of the solution (among the non-dominated solutions, previously identified). The wavelength assignment problem is solved separately, after the bicriteria routing problem.

Let $R = \{N,\ L,\ C,\ T_N\}$ represent the WDM network, where:

- Set of nodes, $N = \{v_1, v_2, \ldots, v_n\}$, $n = \#N$.
- Set of directed arcs, $L = \{l_1, l_2, \ldots, l_m\}$, $m = \#L$.
- Set of wavelengths, $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_W\}$, $W = \#\Lambda$.
- Set of fibres, $F = \{f_1, f_2, \ldots, f_k\}$, $k = \#F$.
- Let $l_i = (v_a, v_b, \bar{o}_{l_i})$, $\bar{o}_{l_i} = (o_{l_i 1}, o_{l_i 2}, \ldots, o_{l_i k})$, $v_a, v_b \in N$.

  If $o_{l_i j} = (1, \bar{a}_j)(j = 1,2,\ldots,k)$, then fibre f_j belongs to arc l_i and contains the wavelengths signalled in $\bar{a}_j$, $\bar{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jW})$, where $a_{ju} = 0,1,2$ $(u = 1,2,\ldots,W)$:

$$a_{ju} = \begin{cases} 0, \text{ if } \lambda_u \text{ does not exist in fibre } f_j\,, \\ 1, \text{ if } \lambda_u \text{ exists and is free in fibre } f_j\,, \\ 2, \text{ if } \lambda_u \text{ exists but is busy in fibre } f_j\,. \end{cases} \quad (1)$$

  If $o_{l_i j} = (0, \bar{a}_j)$ $(j = 1,2,\ldots,k)$, fibre f_j does not belong to arc l_i.

- C is the arc capacity, $C(l_i) = (\bar{n}_{l_i}, \bar{b}_{l_i})$, with $\bar{n}_{l_i} = (n_{l_i 1}, n_{l_i 2}, \ldots, n_{l_i W})$ and $\bar{b}_{li} = (b_{l_i 1}, b_{l_i 2}, \ldots, b_{l_i W})$, where $n_{l_i j}$ is the total number of fibres in arc l_i with wavelength λ_j and $b_{l_i j}$ is the number of fibres where that wavelength is free in arc l_i.

- $T_N(v_i)$ is a table for each node $v_i \in N$ which represents the wavelength conversion capability of the nodes, that is the possibility of transferring the optical signal from one input λ_i to an output λ_j in the node:

$$T_N(v_i) = [t_{uv}], \quad \forall v_i \in N; u, v = 1, 2, \ldots, W, \tag{2}$$

where $t_{uv} = 1(0)$ whether (or not) λ_u can be converted into λ_v, in node v_i.

A *topological path*, p in R, is described by: a source node, a destination node ($v_s, v_t \in N$) and the ordered sequence of nodes and arcs in the path, $p = \langle v_1, l_1, v_2, \ldots, v_{i-1}, l_{i-1}, v_i \rangle$, such that the tail of arc l_k is v_k and the head of l_k is v_{k+1}, for $k = 1, 2, \ldots, i-1$ (all the v_i in p are different).

Besides the ordered sequence of nodes and arcs, a *lightpath* p^λ also comprises the fibre used in each arc and the wavelength on the fibres:

$$p^\lambda = \langle l_c^*, \ldots, l_d^* \rangle = \langle (v_s, v_u, f_i, \lambda_\alpha), \ldots, (v_x, v_t, f_j, \lambda_\beta) \rangle, \tag{3}$$

where $f_i, \ldots, f_j \in F$, $\lambda_\alpha, \ldots, \lambda_\beta \in \Lambda$, represent fibres and wavelengths, respectively.

Note that l_c^* corresponds to $l_c = (v_s, v_u, \bar{o}_{l_c})$ which implies $o_{l_c i} = (1, \bar{a}_i)$ and if $a_{i\alpha} = 1$ then $a_{i\alpha}$ will change from 1 to 2 if p^λ is selected.

With dedicated protection, each connection is supported by two lightpaths (the active lightpath and the protection lightpath), whose topological paths are node disjoint.

2.2. Determination of Node Disjoint Pairs of Topological Paths

Let path $p = \langle v_1, l_1, v_2, \ldots, v_{i-1}, l_{i-1}, v_i \rangle$, be given as an alternate sequence of nodes and arcs from R, such that the tail of l_k is v_k and the head of l_k is v_{k+1}, for $k = 1, 2, \ldots, i-1$ (all the v_i in p are different). Assuming that $N^*(p)$ represents the set of nodes in p, two paths $p = \langle v_1, l_1, v_2, \ldots, v_{i-1}, l_{i-1}, v_i \rangle$ and q are node-disjoint if $\{v_2, \ldots, v_{i-1}\} \cap N^*(q) = \emptyset$.

An algorithm for ranking node disjoint pairs of paths by non-decreasing order of cost, based on an adaption of the MPS algorithm [19], is proposed in [17]. Given an origin-destination node pair, *s*–*t*, the algorithm starts by making a network topology modification (see Fig. 1), where all nodes and links of the graph, (N, L), representing the network topology are duplicated and a new link, of null cost, is added by linking node t to node s' (the duplicate of s): $N' = N \cup \{v_i' : v_i \in V\}$ and $L' = L \cup \{(v_i', v_j') : (v_i, v_j) \in L\} \cup \{(t, s')\}$. In this new augmented graph, (N', L') each path z, from s to t' will correspond to a pair of paths from s to t in (N, L):

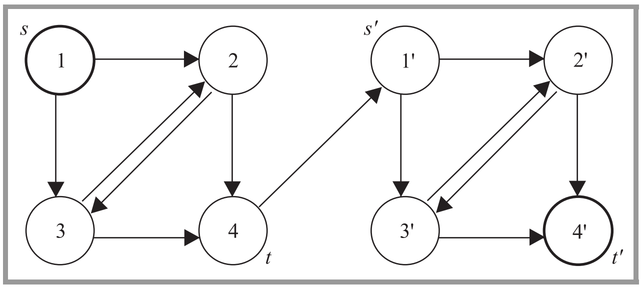


Fig. 1. Topology modification [17].

$$z = p \diamond (t, s') \diamond q \tag{4}$$

where p is a path from s to t in (N, L) and q is a path from s' to t' in (N', L').

Finally, the adapted version of MPS is used for ranking by non-decreasing order of cost the paths z, such that p and q are node disjoint. Let the set of paths from a source node s to a destination node t in (N, L) be $\mathscr{P}_{st}$. Note that each path z from s to t' in (N', L') is given by (4), with $p \in \mathscr{P}_{st}$ and $q \in \mathscr{P}'_{s't'}$.

2.3. Bicriteria Approach

Having in mind a bicriteria optimization model, we consider two additive objective functions for the active and the protection path – the first one is the sum of the inverse of the available bandwidth in the links of each path and the second is the number of links (or hop count) of the paths. The duplicated links in the augmented graph, (N', L') also have the same costs and the two costs of link (t, s') are null. The first objective function, $c_1(z)$ is related to the bandwidth usage in the links of the path z and is expressed in the inverse of the available bandwidth in the links:

$$\min_{z \in D} \left\{ c_1(z) = \sum_{l \in z} \frac{1}{b_l^T} = \sum_{l \in p} \frac{1}{b_l^T} + \sum_{l \in q} \frac{1}{b_l^T} \right\}, \tag{5}$$

where D is the set of topological paths for the origin–destination node pair (s, t') and b_l^T is the total available capacity in link l, in terms of available wavelengths. This criterion seeks to avoid already congested links, favoring a balanced distribution of traffic throughout the network, and hence decreasing the blocking probability and therefore increased the expected revenue. The same criterion was used in the model without protection analyzed in the related paper [15]. The values of the available bandwidths b_l^T to be used in each instance of the resolution of the bi-objective optimization problem are calculated from the vector $\bar{b}_l$ in $C(l)$:

$$b_l^T = \sum_{j=1}^{W} b_{lj}, \quad \forall l \in L. \tag{6}$$

The second objective consists of minimizing the sum of the number of links of the two paths, $h(p) + h(q)$, seeking to avoid bandwidth waste, hence favoring global efficiency in the use of network resources:

$$\min_{z \in D} \left\{ c_2(z) = h(p) + h(q) \right\}. \tag{7}$$

Note that in many cases there is no feasible solution which optimizes the two objective functions, $c_1(z)$ and $c_2(z)$, simultaneously. A certain amount of conflict is therefore expected between c_1 and c_2, and no optimal solution (in most cases) will exist for this problem. Therefore the candidate solutions to the topological RWA multicriteria model are topological paths which are non-dominated solutions to the bi-objective problem:

$$(\mathscr{P}) \quad \begin{cases} \min_{z \in D_T} c_1(z) \\ \min_{z \in D_T} c_2(z) \end{cases} . \tag{8}$$

The set of admissible solutions, D_T, consists of all topological paths between the source-destination node pair (s,t') in (N',L') which correspond to node disjoint paths pairs (p,q) in (N,L) and to *viable lightpaths* (p^λ, q^λ), that is, lightpaths with the same arcs as p and q and with a free and usable wavelength (according to T_N) in every arc. The topological paths in these conditions (elements of D_T) will be designated as *viable topological paths*, for the given origin-destination node pair. Firstly, for obtaining D_T, the free wavelengths in each arc will have to be identified, taking into account the wavelength conversion capabilities specified by T_N, then the set of viable node disjoint paths pairs (p^λ, q^λ) for the origin-destination node pair becomes implicitly defined.

This model was extended to bidirectional connections between nodes s and t by considering a bidirectional lightpath $z^\lambda = (z^\lambda_{st'}, z^\lambda_{t's})$ supported by a bidirectional topological path $z = (z_{st'}, z_{t's})$ which is a pair of symmetrical topological paths in (N',L'). In this case the set D^b_T of feasible solutions to the bicriteria model will be the set of viable bidirectional topological paths z, i.e., characterized by the fact that both (unidirectional) topological paths $z_{st'}$ and $z_{t's}$ are viable. Therefore the bi-objective bidirectional routing optimization problem is formulated as:

$$\min_{p \in D^b_T} \left\{ c_1(z) = \sum_{l \in p_{st}} \frac{1}{b^T_l} + \sum_{l \in q_{s't'}} \frac{1}{b^T_l} + \sum_{l \in p_{ts}} \frac{1}{b^T_l} + \sum_{l \in q_{t's'}} \frac{1}{b^T_l} \right\}, \tag{9}$$

$$\min_{p \in D^b_T} \left\{ c_2(z) = h(p_{st}) + h(q_{s't'}) + h(p_{ts}) + h(q_{t's'}) \right\} . \tag{10}$$

We will assume the most common situation in real networks where the two paths $z_{st'}, z_{t's}$ are topologically symmetrical, thence $c_2(z) = 2[h(p_{st}) + h(q_{s't'})]$. Note that this does not imply that the wavelengths used in the two opposite directions are necessarily symmetrical.

2.4. Resolution Method

The addressed problem is: given a source-destination pair of nodes, $s-t$, find a pair (p,q) of node disjoint paths which minimises $c_i(z) = c_i(p) + c_i(q)$, $i = 1,2$.

As in [17], we will say that, given two node disjoint path pairs (p_j, q_j) $(j = 1,2)$ from s to t in R, pair (p_1, q_1) dominates (p_2, q_2), denoted by $(p_1,q_1)_D(p_2,q_2)$, if and only if $c_i(p_1) + c_i(q_1) \le c_i(p_2) + c_i(q_2)$ $(i = 1,2)$ and at least one of the inequalities is strict. A node disjoint path pair (p,q) is a non dominated solution if no other feasible node disjoint path pair dominates it.

The aim of the resolution procedure is to find a *good* compromise node disjoint path pair from the set of non-dominated solutions, according to certain criteria, previously defined. Secondly, one must note that path calculation and selection have to be fully automated, having in mind the nature of a telecommunication network routing mechanism, so that an interactive decision approach is precluded.

Topological paths $z = p \diamond (t,s') \diamond q$ that are candidate solutions of the problem are generated in the modified graph according to the algorithm in [17], using as path cost a convex combination of the two objective functions $f(z) = \alpha c_1(z) + (1-\alpha) c_2(z)$ – recall that the arc (t,s') has null cost in both metrics. The value of α is not relevant and only defines the order by which solutions will be obtained by the algorithm for ranking node disjoint pairs of paths by cost f. Every generated solution will have to be evaluated to determine if it can correspond to a viable lightpaths and then a dominance test is used to determine whether or not it is non-dominated with respect to all the previously generated solutions. Only viable lightpaths which are non-dominated solutions will be stored.

The selection of the final solution follows a procedure perfectly analogous to the one used for the bicriteria model without protection [14], [15]. It is based on the definition of preference thresholds for both functions in the form of requested and acceptable values, and on a reference point like approach (see detailed description in [16]). These thresholds enable the specification of priority regions in the objective function's space.

Let $z^{c_1} = p^{c_1} \diamond (t,s') \diamond q^{c_1}$ be the shortest path with respect to the first objective function, and z^{c_2} the shortest path with respect to the second objective function (computed by solving the associated shortest path problems). This leads to the ideal solution, $\mathscr{O}$, in the objective functions' space:

$$z^{c_1} = \arg\min_{z \in D_T} \{c_1(z)\} , \tag{11}$$

$$z^{c_2} = \arg\min_{z \in D_T} \{c_2(z)\} . \tag{12}$$

The objective functions space, where non-dominated solutions will be searched, is defined by the points (c_{1m}, c_{2M}) and (c_{1M}, c_{2m}):

$$c_{1m} = c_1(z^{c_1}) = c_1(p^{c_1}) + c_1(q^{c_1}) , \tag{13}$$

$$c_{2M} = c_2(z^{c_1}) = c_2(p^{c_1}) + c_2(q^{c_1}) , \tag{14}$$

$$c_{1M} = c_1(z^{c_2}) = c_1(p^{c_2}) + c_1(q^{c_2}) , \tag{15}$$

$$c_{2m} = c_2(z^{c_2}) = c_2(p^{c_2}) + c_2(q^{c_2}) . \tag{16}$$

The preference thresholds $c_{1\text{req}}$, $c_{2\text{req}}$ (requested values) and $c_{1\text{acc}}$, $c_{2\text{acc}}$ (acceptable values) that circumscribe the priority regions are defined (taking into account the discrete nature of $c_2(z)$) by the following expressions:

$$c_{1\text{acc}} = c_{1M} , \tag{17}$$

$$c_{2\text{acc}} = c_{2M} , \tag{18}$$

$$c_{1\text{req}} = \frac{c_{1m} + c_{1M}}{2} , \tag{19}$$

$$c_{2\text{req}} = \left\lfloor \frac{c_{2m} + c_{2M}}{2} \right\rfloor , \tag{20}$$

which result in four priority regions in the objective functions' space (as in [15]).

The selection of the final solution, when there is more than one non-dominated solution in a region S, uses a reference point based procedure of the type proposed in [20]. In the present context we used a weighted Chebyshev metric [18] proportional to the size of the "rectangle" S:

$$\min_{z\in S}\max_{i=1,2}\left\{w_i|c_i(z)-\underline{c}_i|\right\}, \tag{21}$$

where $(\underline{c}_1,\underline{c}_2)$ is the reference point, which is chosen as the left down corner of region S; the right upper corner is given by $(\bar{c}_1,\bar{c}_2)$, and the weights w_i $(i=1,2)$ are:

$$w_i=\frac{1}{|\bar{c}_i-\underline{c}_i|}. \tag{22}$$

Details of this selection procedure can be seen in [14], [20].

This resolution method seeks to make the most of the very great efficiency of the used shortest path ranking algorithm [21], [17] (used to calculate candidate solutions) and the inherent superiority of the use of a reference point-based procedure, as a solution selection mechanism. Note that the automated nature of the routing mechanism (with protection) requires a solution in a very short time period. The final stage of the resolution method is the selection of the wavelengths along the arcs of the selected path, described in the next subsection.

The proposed resolution approach can be applied straightforwardly to the calculation and selection of bidirectional lightpaths, with the necessary adaptation to the objective functions, according to the definitions in (9) and (10).

2.5. Wavelength Assignment Heuristic

After the selection of the pair of topological node disjoint paths (unidirectional or bidirectional), the second stage is the assignment of wavelengths (and corresponding fibres) along the links of the paths, hence completing the lightpaths specification. Wavelength selection seeks to maximise the *wavelength bottleneck bandwidth*, $b_j(p)$ $(\lambda_j\in\Lambda)$:

$$\max_{\lambda_j\in\Lambda}\left\{b_j(p)=\min_{l\in p\wedge b_{lj}>0} b_{lj}\right\}, \quad (p\in D_T). \tag{23}$$

This procedure corresponds to the choice of the least loaded wavelength (LL) along the arcs of the path p. Note that if all the nodes of the network enable full wavelength conversion, once a viable topological path is chosen, the choice of the wavelength(s) to be used is irrelevant in terms of network performance. If the nodes have no conversion capability the proposed criterion of wavelength selection is known in the literature (see, e.g., [4]) to give good results. In any case it is also known that in these cases the critical factor in terms of network performance is the selection of topological paths, the choice of wavelength having a minor impact.

In the present model this choice of wavelength will correspond to specify λ_{j^*} in arc l^*:

$$b_{l^*j^*}=\max_{\lambda_j\in\Lambda}\left\{b_j(p)=\min_{l\in p\wedge b_{lj}>0} b_{lj}\right\}: \begin{array}{l}\exists \text{ viable } p^{\lambda} \text{ which}\\ \text{uses } \lambda_{j^*} \text{ in } l^*\in p.\end{array} \tag{24}$$

Further details and an illustrative example of this selection heuristic can be seen in [14].

The same procedure is used for wavelength and fibre selection along the links of the node disjoint path q.

For bidirectional connections, once a non-dominated solution $z\in D_T^b$ has been selected, the wavelengths (and fibres) to be used along $z_{st'}$ and $z_{t's}$ are chosen applying the same procedure to each path. Note that the chosen wavelength(s) in each path can be different.

3. Performance Analysis of the Bicriteria Model with Protection

Extensive simulations with the model were made on several typical WDM networks found in literature. This section presents the simulation results in four of such networks, namely, the NSFNET [22] (see Fig. 2), the Pan-European network COST 266BT [22] (Fig. 3), a typical core network presented in [23] – Kodialam network (KL) (Fig. 4), and a typical network provider network presented in [24] – ISP network (Fig. 5). Table 1 summarizes the main characteristics of these networks. All the networks were dimensioned for about one thousand bidirectional lightpaths (1084 for

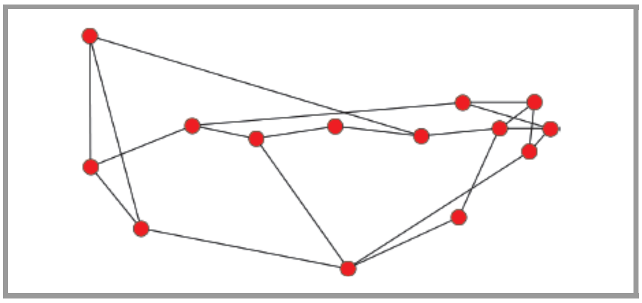


Fig. 2. NSFNET network (14 nodes and 21 links) [22].

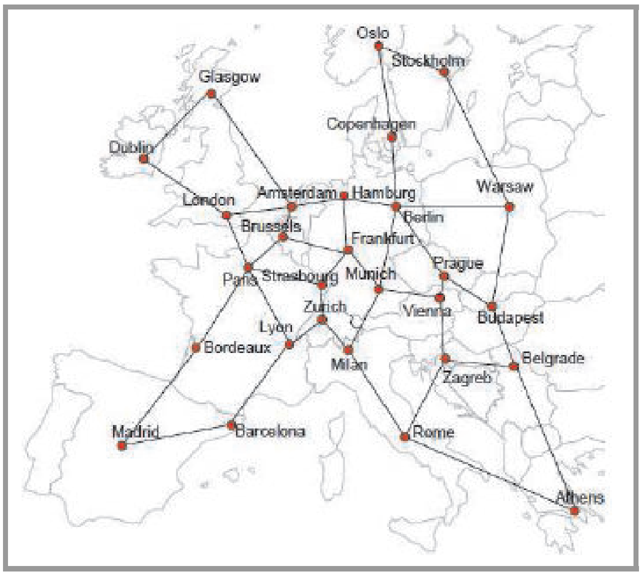


Fig. 3. COST 266BT Pan-European network (28 nodes and 41 links) [22].

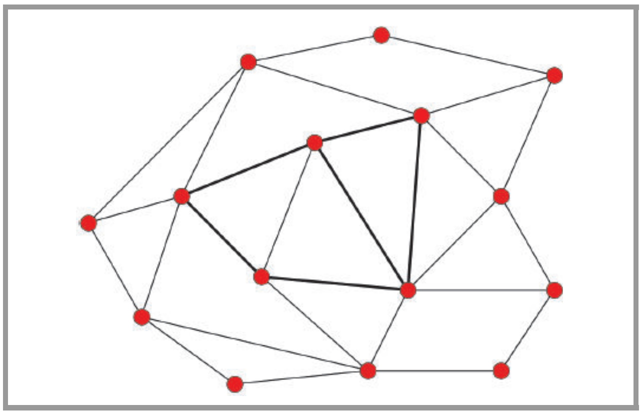


Fig. 4. KL network (15 nodes and 28 links) [23].

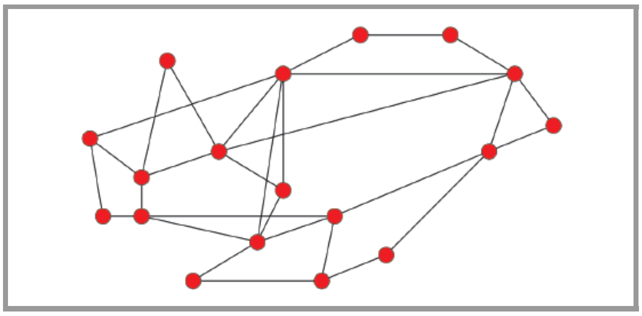


Fig. 5. ISP network (18 nodes and 30 links) [24].

NSFNET, 1008 for COST 266BT, 1050 for KL network, and 918 for ISP network) and each fibre has 16 wavelengths.

Table 1
Networks characteristics

Network	Number of		Nodal
	nodes	links	degree
NSFNET [22]	14	21	3.00
COST266BT [22]	28	41	2.93
KL [23]	15	28	3.73
ISP [24]	18	30	3.33

Two different scenarios of conversion capability were considered in simulations: all nodes without conversion capability (first scenario) and only five nodes with total conversion capability (central nodes were chosen with this capability) – second scenario.

Simulations were run up to 1200 requests (incremental traffic) in two different cases: with 100% bidirectional requests and with 5% unidirectional requests (usually, most of the connection requests for lightpaths are bidirectional).

The simulations showed that the performance variation due to presence of five nodes with total conversion capability is negligible. Therefore, from now on, we only present the scenario without conversion.

For performance assessment purposes, the results in several relevant network performance measures obtained with the bicriteria model (BiC) will be compared with the corresponding results of the single objective formulations, namely, the first objective function related with the bandwidth usage (SP_c1), and the second objective function, concerning hop count (SP_c2).

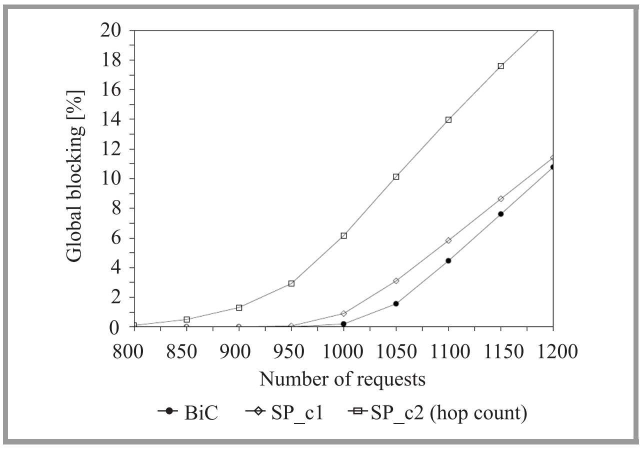


Fig. 6. Global blocking – NSFNET network.

Figure 6 shows that the blocking probability in the NSFNET for the BiC model has a value significantly lower than in the SP_c2 model. It is also lower than the blocking

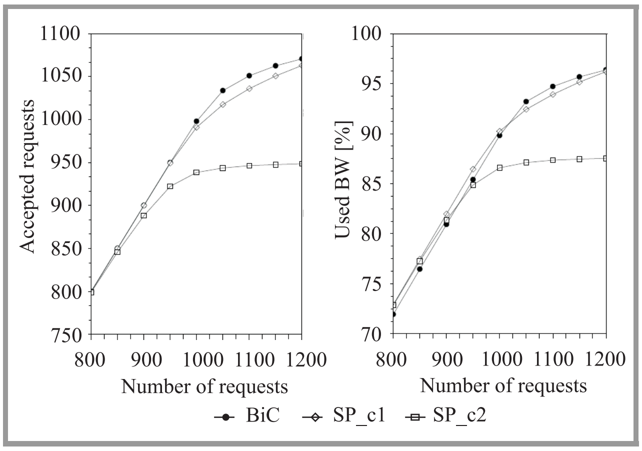


Fig. 7. Accepted requests versus used bandwidth – NSFNET network.

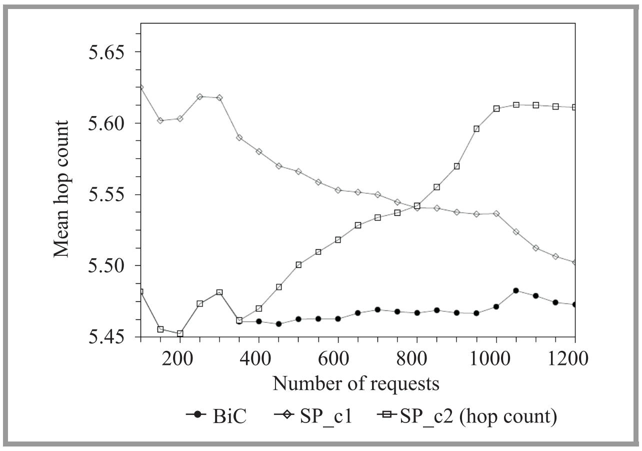


Fig. 8. Mean hop count – NSFNET network.

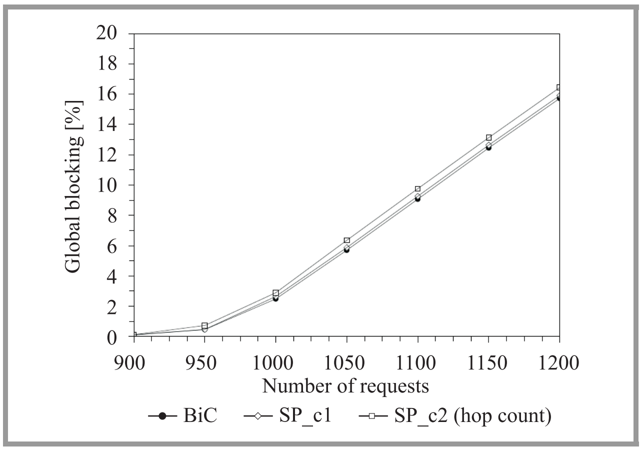


Fig. 9. Global blocking – COST 266BT network.

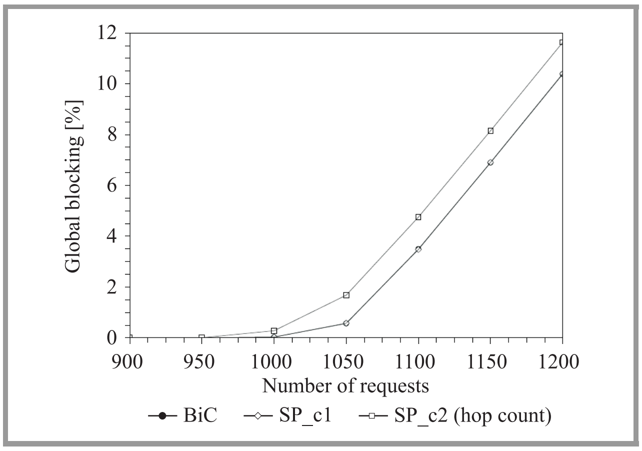


Fig. 10. Global blocking – KL network.

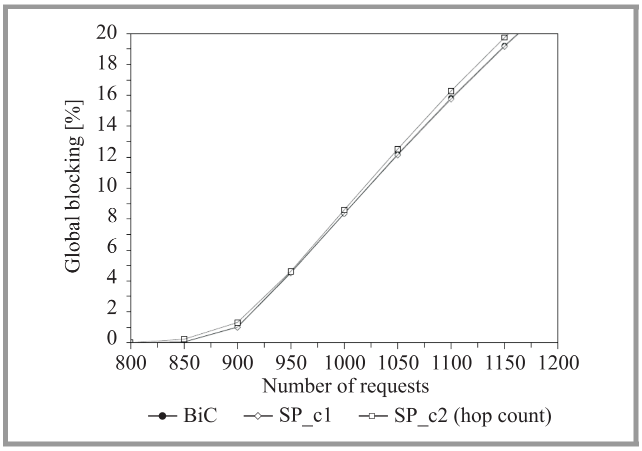


Fig. 11. Global blocking – ISP network.

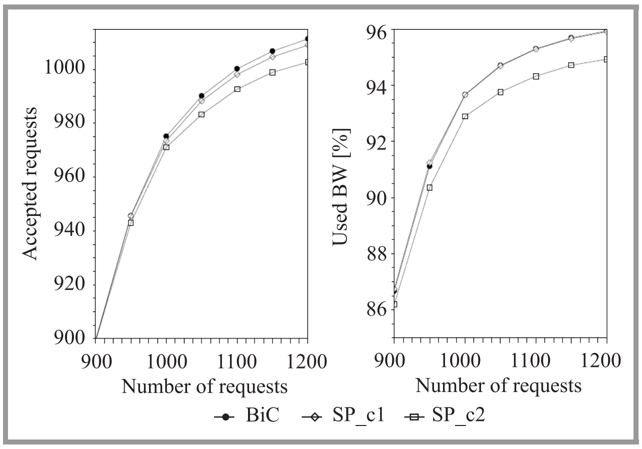


Fig. 12. Accepted requests versus used bandwidth – COST 266BT network.

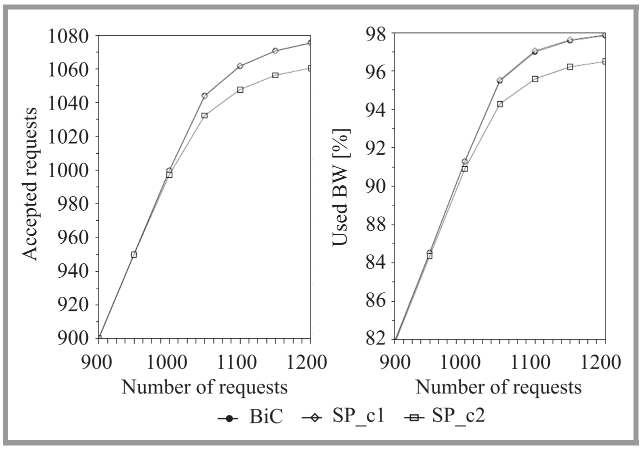


Fig. 13. Accepted requests versus used bandwidth – KL network.

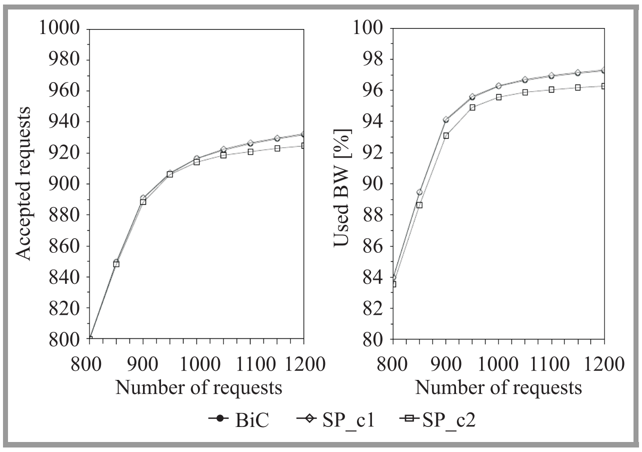


Fig. 14. Accepted requests versus used bandwidth – ISP network.

probability observed in SP_c1, although the difference is smaller. The BiC and SP_c1 models do not exhibit blocking until 950 connection requests. SP_c2 performs worse, as blocking appears for approximately 850 connection requests.

As it can be seen in Fig. 7, for moderate traffic loads (up to 1000 requests), although the number of accepted connections is higher in BiC, it uses less bandwidth than SP_c1. Above 1000 requests, the SP_c1 model requires less bandwidth than BiC, but this happens because SP_c1 accept less requests. The lowest average number of hops per connection (see Fig. 8) also shows the efficiency of the BiC formulation – BiC normally chooses shorter paths.

Figure 7 shows that the used network bandwidth in BiC and SP_c1 exceeds 95%, above approximately 1050 accepted requests, a value similar to the number of connections for which the network was dimensioned (1084 in NSFNET network).

Although not shown in the figures, the topologies with five nodes with complete conversion capability offers a negligible performance improvement. The results obtained when 5% of the requests were unidirectional are similar to the ones with 100% bidirectional connections.

The global blocking probability for the COST 266BT[1], KL and ISP networks with protection is shown in Figs. 9–11. Figures 12–14 show the number of accepted requests and the used bandwidth for the same topologies (Figs. 9, 10, 12, and 13 only show the results above 900 connection requests because, below this value, the blocking probability for KL and COST 266BT networks is almost zero).

Regarding the blocking probability on these networks, BiC model clearly exhibits a better performance than SP_c2. On the COST 266BT network the blocking in BiC model is only slightly lower than in SP_c1. Figure 12 shows that BiC and SP_c1 use the same amount of bandwidth but the number of accepted lightpaths in the BiC model is slightly larger. But, contrary to the results obtained without protection [15], the BiC and SP_c1 approaches applied to KL and ISP networks with protection have roughly the same performance. So the BiC model for dedicated path protection has not always a better performance than the SP_c1 – in some topologies, the single criterion model based on the bandwidth usage in the links of the path has a global blocking probability similar to the bicriteria model.

Regarding the traffic distribution capability of the three models, Fig. 15 shows the number of arcs with less than 10% free bandwidth in the NSFNET network. Until 1000 requests in the NSFNET network (the only one where BiC is clearly better than SP models) the BiC model provides a lower number of arcs with less than 10% free bandwidth (Fig. 15), although it has a slightly higher number of accepted requests. For COST 266BT, KL and ISP networks this measure has a similar behavior in BiC and SP_c1 models.

Concerning CPU times in an AMD 64X2 processor at 2.4 GHz, they are very low. In NSFNET the CPU time is approximately 0.25 ms for single objective formulations and 0.5 ms for BiC (Fig. 16). Note that these CPU times are roughly twice those obtained in the model without protection (see [15]). In COST 266BT network the BiC uses less than 1 ms below 900 requests while single objective approaches use about 0.5 ms (see Fig. 17). When the number of requests exceeds 900 the CPU time grows up to 2.4 ms in BiC, 2.1 ms in SP_c1 and up to 1 ms in SP_c2. In the KL network up to 1000 requests, SP_c1 and SP_c2 use about 0.27 ms per connection request, while BiC uses 0.5 ms (roughly twice the CPU time obtained without protection [15]). In the ISP network the CPU times are slightly higher, about 0.3 ms for the SP_c1 and SP_c2 approaches and 0.5 ms for BiC, until 900 requests. The CPU time

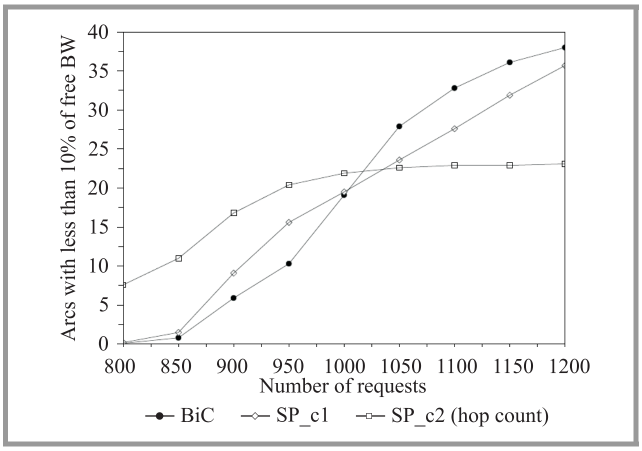


Fig. 15. Arcs with less than 10% of free BW – NSFNET network.

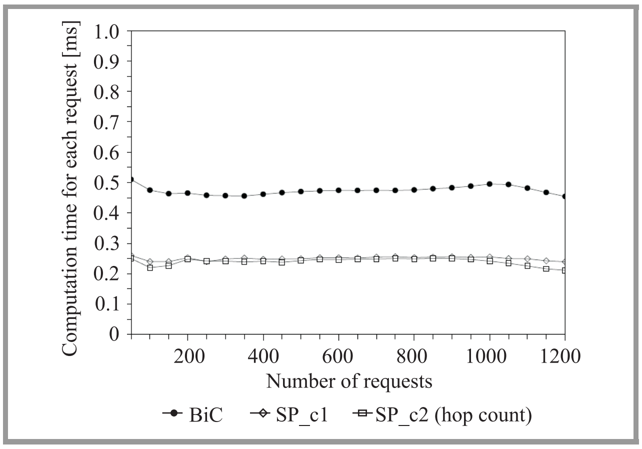


Fig. 16. Computation time for each request – NSFNET network.

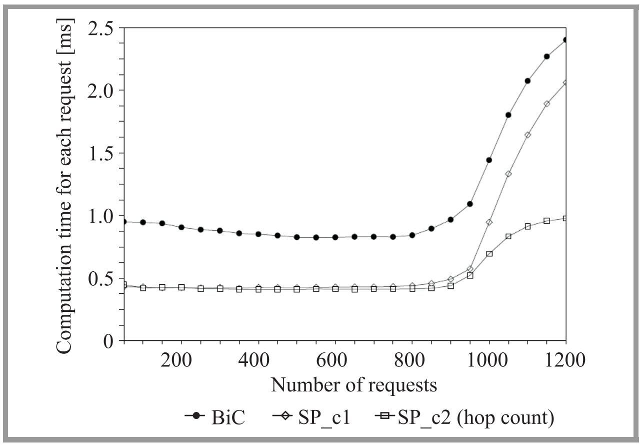


Fig. 17. Computation time for each request – COST 266BT network.

[1] Comparing the results for global blocking probability in the COST 266BT network with those presented in [16], apparently for the same network, a significant performance improvement can be verified. This is due to a different network dimensioning. The simulations in [16] use the network dimensioning presented in [22], which results in a total of 1066 fibres of 16 wavelengths each, while here we use a dimensioning method in line with the routing scheme. The total resources are only slightly different – 1094 fibres – but their distribution in the 41 bidirectional links of the network is substantially different.

increase, verified in COST 266BT, KL, and ISP networks coincides with the starting of visible blocking.

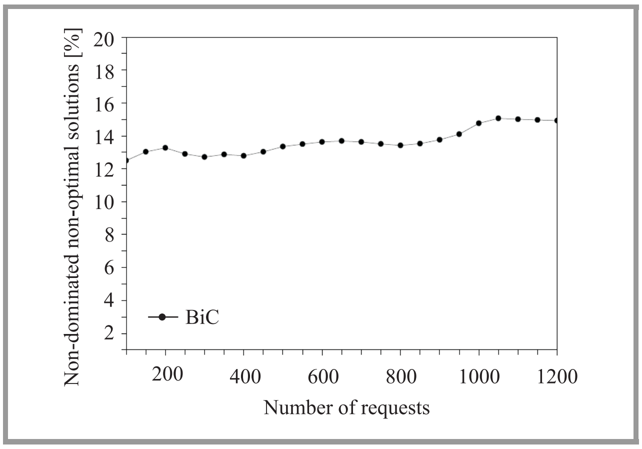


Fig. 18. Non-dominated non-optimal solutions – NSFNET network.

To assess the degree of conflict between the two objective functions involved in the bicriteria model, the number of accepted requests with optimal solution was measured. Figure 18 shows the number of requests without an optimal solution in the NSFNET network. This number of non-dominated solutions is relatively low, which indicates a relatively low degree of conflict between the functions c_1 and c_2, but, at least in some networks/topologies, the bicriteria model exceeds the performance of the single criteria approaches.

4. Conclusions

The routing and wavelength assignment problem in WDM networks involves multiple objectives and constraints, so, multicriteria approaches like the one analyzed in this paper enable an explicit representation of the different performance objectives and the addressing, in a mathematically consistent manner, of the trade offs among the various criteria.

A bicriteria model for obtaining a topological pair of node-disjoint paths unidirectional or symmetric bidirectional for each connection request in WDM networks was analyzed in terms of relevant network performance metrics. The optimization model considers two criteria – one concerning the bandwidth usage in the links of the network and the other the number of links of the paths. All the non-dominated solutions are identified using an efficient k-shortest path algorithm, applied to a modified topology. The automated selection of final solution uses preference thresholds defined in the objective function's space, combined with a Chebyshev distance to a reference point. Having obtained the "best" non-dominated topological path pair, a heuristic procedure was then used to assign wavelengths to the links of the paths.

Several benchmark networks were used to perform extensive network performance assessment of this bicriteria model, considering a comparison with the results of the two single criterion approaches corresponding to each of the criteria used in the BiC model. The impact of having five nodes with wavelength conversion capability was negligible in the simulated situations. The BiC model leads to a performance better than the monocriteria model SP_c2 (based on the hop count metric).

Regarding the comparison between BiC and SP_c1 approaches, only in one of the simulated networks the performance of BiC was clearly better than SP_c1. This happens in the smaller network (NSFNET). In all other cases, and contrary to what happens in the model without protection, with dedicated protection the BiC and SP_c1 approachs have similar performance in some cases. So the bicriteria model (with these two criteria) for dedicated path protection does not seem to provide additional benefits in all networks topologies as compared to the single criterion model based on link usage costs.

Although the BiC model uses more CPU time than the monocriteria approaches its values are quite low, even when the networks are congested.

Acknowledgements

This study was partially supported by PT Inovação and by programme POSC of the EC programme cosponsored by national funds.

C. Simões also thanks the FCT and IPV by the grant of the Ph.D. scholarship SFRH/BD/49739/2009.

References

[1] J. Zhang, K. Zhu, L. H. Sahasrabuddhe, S. J. B. Yoo, and B. Mukherjee, "On the study of routing and wavelength assignment approaches for survivable wavelength routed WDM mesh networks", *Opt. Netw. Mag.*, vol. 4, pp. 16–28, 2003.

[2] J. Zhang and B. Mukherjee, "A review of fault management in WDM mesh networks: Basic concepts and research challenges", *IEEE Netw.*, vol. 18, pp. 41–48, 2004.

[3] C. S. Ou, H. Zang, N. K. Singhal, K. Zhu, L. H. Sahasrabuddhe, R. A. MacDonald, and B. Mukherjee, "Subpath protection for scalability and fast recovery in optical wdm mesh networks", *IEEE J. Sel. Areas Commun.*, vol. 22, pp. 1859–1875, 2004.

[4] H. Zang, J. P. Jue, and B. Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks", *Opt. Netw. Mag.*, vol. 1, pp. 47–60, 2000.

[5] X. Chu, J. Liu, and Z. Zhang, "Analysis of sparse-partial wavelength conversion in wavelength-routed WDM networks", in *Proc. IEEE Conf. INFOCOM 2004*, 2004, Hong Kong, China, vol. 2, pp. 1363–1371.

[6] I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: an approach to high bandwidth optical WAN's", *IEEE Trans. Commun.*, vol. 40, pp. 1171–1182, 1992.

[7] D. Xu, Y. Xiong, C. Qiao, and G. Li, "Failure protection in layered networks with shared risk link groups", *IEEE Netw.*, vol. 18, pp. 36–41, 2004.

[8] C. Xin, Y. Ye, S. S. Dixit, and C. Qiao, "A joint lightpath routing approach in survivable optical networks", *Opt. Netw. Mag.*, vol. 3, pp. 13–20, 2002.

[9] A. Sen, B. Hao, B. H. Shen, and S. Bandyopadhyay, "Survivability of lightwave networks – path lengths in WDM protection scheme", *J. High Speed Netw.*, vol. 10, no. 4, pp. 303–315, 2001.

[10] R. Bhandari, *Survivable Networks: Algorithms for Diverse Routing*. Norwell: Kluwer, 1998.

[11] J. W. Suurballe and R. E. Tarjan, "A quick method for finding shortest pairs of disjoint paths", *Networks*, vol. 14, no. 2, pp. 325–336, 1984.

[12] J. Clímaco and J. Craveirinha, "Multicriteria analysis in telecommunication planning and design – problems and issues", in *Multiple Criteria Decision Analysis: State of the Art Surveys*. International Series in Operations Research & Management Science, vol. 78, pp. 899–951. Heidelberg: Springer, 2005.

[13] J. C. N. Clímaco, J. M. F. Craveirinha, and M. M. B. Pascoal, "Multicriteria routing models in telecommunication networks – overview and a case study", in *Advances in Multiple Criteria Decision Making and Human Systems Management: Knowledge and Wisdom*. Amsterdam: IOS Press, 2007, pp. 17–46.

[14] T. Gomes, J. Craveirinha, J. Clímaco, and C. Simões, "A bicriteria routing model for multi-fibre WDM networks", *Phot. Netw. Commun.*, vol. 18, pp. 287–299, 2009.

[15] C. Simões, T. Gomes, J. Craveirinha, and J. Clímaco, "Performance analysis of a bi-objective model for routing and wavelength assignment in WDM networks", *J. Telecommun. Inform. Technol.*, no. 3, pp. , 2010.

[16] T. Gomes, C. Simões, J. Craveirinha, and J. Clímaco, "A bi-objective model for routing and wavelength assignment in resilient WDM networks", *Saf. Rel. Risk Anal. Theory, Meth. Appl.*, vol. 4, pp. 2627–2634, 2008.

[17] J. C. N. Clímaco and M. M. B. Pascoal, "Finding non-dominated bicriteria shortest pairs of disjoint simple paths", *Comput. Oper. Res.*, vol. 36, pp. 2892–2898, 2009.

[18] A. P. Wierzbicki, "The use of reference objectives in multiobjective optimization", *Theor. Appl. Proc. Lect. Notes Econom. Math. Syst.*, vol. 177, pp. 468–487, 1980.

[19] E. Martins, M. Pascoal, and J. Santos, "Deviation algorithms for ranking shortest paths", *Int. J. Found. Comput. Sci.*, vol. 10, no. 3, pp. 247–263, 1999.

[20] J. C. N. Clímaco, J. M. F. Craveirinha, and M. M. B. Pascoal, "An automated reference point-like approach for multicriteria shortest path problems", *J. Syst. Sci. Syst. Eng.*, vol. 15, pp. 314–329, 2006.

[21] E. Martins, M. Pascoal, and J. Santos, "An algorithm for ranking loopless paths", Tech. Rep. 99/007, CISUC, 1999 [Online]. Available: http://www.mat.uc.pt/˜marta/Publicacoes/mps2.ps

[22] A. Betker, C. Gerlach, M. Jäger, M. Barry, S. Bodamer, J. Späth, C. M. Gauger, and M. Köhn, "Reference transport network scenarios", Tech. Rep., MultiTeraNet Report, July 2003.

[23] M. Kodialam and T. V. Lakshman, "Dynamic routing of bandwidth guaranteed tunnels with restoration", in *Proc. IEEE Conf. INFOCOM 2000*, Tel-Aviv, Israel, 2000, vol. 2, pp. 902–911.

[24] G. Apostolopoulos, R. Guérin, S. Kamat, and S. K. Tripathi, "Quality of service based routing: a performance perspective", in *Proc. ACM SIGCOMM'98 Conf.*, Vancouver, Canada, 1998, pp. 17–28.

Carlos Simões, João Clímaco – for biographies, see this issue, p. 24.

Teresa Gomes, José Craveirinha – for biographies, see this issue, p. 12.

Paper

A Survey of Multi-Objective Deployment in Wireless Sensor Networks

Michał Marks

Abstract—The major challenge in designing wireless sensor networks (WSNs) is to find tradeoff between the desired and contrary requirements for the lifetime, coverage or cost while coping with the computation, energy and communication constraints. This paper examines the optimal placement of nodes for a WSN. It is impossible to consider the deployment of the nodes separately from WSNs applications. We highlight the properties of WSNs applications that determine the placement problem. We identify and enumerate the various objectives that should be considered. The paper provides an overview and concentrates on multi-objective strategies, their assumptions, optimization problem formulation and results.

Keywords—coverage, lifetime, placement, positioning, wireless sensor network.

1. Introduction

In recent years, with advance in wireless communication technology, sensing technology, micro-electronics technology and embedded system, wireless sensor networks can be used for a wide variety of applications and systems with vastly varying requirements and characteristics, such as environmental monitoring, disaster management, factory automation, health care or military. Typical sensor network consists of a large number of spatially distributed autonomous sensor devices. Nodes networked through wireless must gather local data and communicate with other nodes.

A wireless sensor network (WSN) design is influenced by many factors such as transmission errors, network topology and power consumption. Consequently, developing a WSN application introduces several implementation challenges. This paper describes one of the most fundamental issue in WSN designing – the deployment problem. This specific problem has different appellations in the literature, e.g., placement, layout, coverage or positioning problem in WSNs. The term positioning seems to be more general, so we propose a taxonomy illustrated in Fig. 1. On the left is localization – its aim is to locate where the nodes are placed. On the right is deployment (placement) – its aim is to determine where the nodes should be placed. In the vast majority of deployment problems the coverage is considered, but this is not necessary and depends on the application. More details about the applications and its properties can be found in Section 2.

In this paper we concentrate on optimal node placement. This is one of the most important design step to selectively decide the locations of the sensors to optimize the desirable objectives, e.g., maximize the covered area or minimize the energy use. Fundamental questions in this case include [1]:

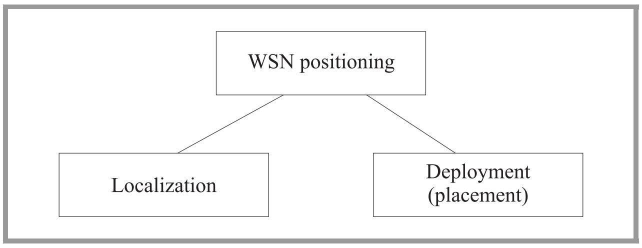


Fig. 1. A taxonomy for positioning in WSN.

- How many sensor nodes are needed to meet the overall system objectives?
- For a given network with a certain number of sensor nodes, how do we precisely deploy these nodes in order to optimize network performance?
- When data sources change or some part of the network malfunctions, how do we adjust the network topology and sensor deployment?

2. Wireless Sensor Network Applications and Properties

In the past, a number of early, mostly US-based research projects established a de facto definition of a wireless sensor network as a large-scale, wireless, ad hoc, multi-hop, unpartitioned network of largely homogenous, tiny, resource-constrained, mostly immobile sensor nodes that would be randomly deployed in the area of interest [2].

More recently WSNs are used in a huge variety of scenarios. Such diversity translates into different requirements and the above definition of a wireless sensor network does not necessarily apply for those scenarios. The knowledge about sensor networks is evolving in many different directions. Of course we still have a classical sensor networks but now we can also distinguish a mobile sensor networks [3], wireless sensor and actuator networks [4], wireless multimedia sensor networks [5] and many others. This coarse-grained division cannot be treated as a classification of sensor networks. It illustrates only some emerging trends which enhances diversity in WSNs. In many applications a network is small-scale with a few dozens of nodes, some nodes are mobile, they are not homogeneous etc. This diversity can be considered in many different dimensions. Römer and Mattern [2] propose over ten properties characterizing existing WSN applications such as size, mobility, heterogeneity, communication modality etc. Another taxonomy can be found in Mottola and Picco survey [6], illustrated in Fig. 2.

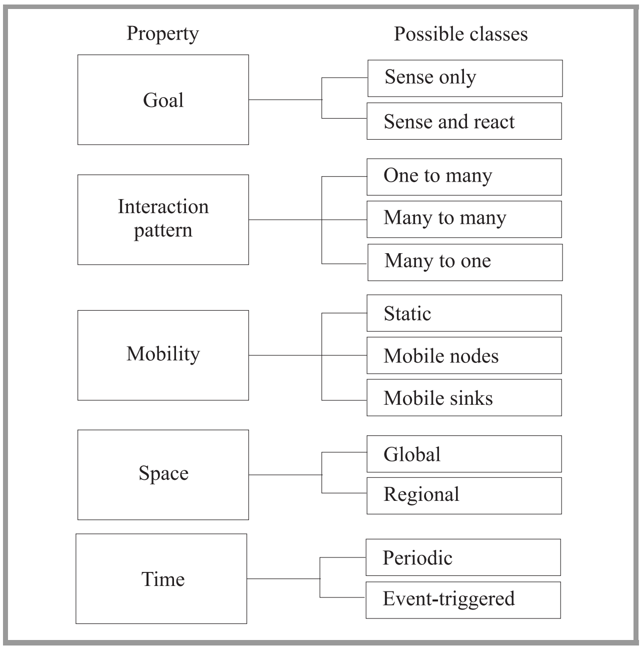


Fig. 2. A taxonomy of WSN applications by [6].

Goal. In the majority of WSNs applications, especially the early ones, the goal is to gather environmental data for later analysis (*sense-only*). This can be done by a field of sensor-equipped nodes which sends their data, possibly along multiple hops, to a single base station that centrally collects the readings. However now we can distinguish also applications where some WSN nodes are equipped with actuators. In WSANs, the roles of sensor and actor nodes are to collect data from the environment and perform appropriate actions based on this collected data, respectively *sense and react* [4].

Interaction pattern. Another key property is interaction pattern between the network nodes – the way how the network nodes exchange information with each other, which is somehow affected also by the application goal they are trying to accomplish. The most popular interaction pattern is *many-to-one*, where data is send from all nodes in the network to a central collection point. Nevertheless, *one-to-many* and *many-to-many* interactions can also be found. The former are important when it is necessary to send configuration commands (e.g., a change in the sampling frequency or in the set of sensors active) to the nodes in the network. The latter is typical of scenarios where multiple data sinks are present, a situation often manifest in sense-and-react scenarios.

Mobility. This is probably the most noticeable property. Sensor nodes may change their location after initial deployment. Mobility may apply to all nodes within a network or only to subsets of nodes. Mottola and Picco [6] distinguish three classes *static*, *mobile nodes* and *mobile sinks*. Roughly the same classes are described in [2], but Römer indicates also some other aspects of mobility – shown in Fig. 3. Mobility can result from environmental influences such as wind or water, sensor nodes may be attached to or carried by mobile entities – *passive mobility*. Sensor nodes may possess automotive capabilities – *active mobility*. The degree of mobility may also vary from *occasional* movement with long periods of immobility in between, to *constant* travel.

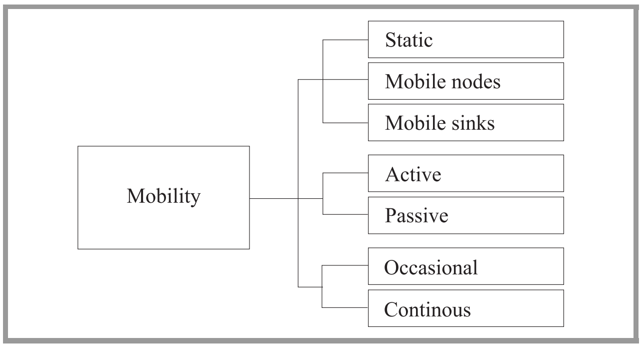


Fig. 3. Extended mobility taxonomy.

Space. Different applications may require the distributed processing spreading different portions of the physical space. The space can be *global* where the processing involves in principle the whole network, most likely because the phenomena of interest span the whole geographical area where the WSN is deployed or *regional* where the majority of the processing occurs only within some limited area of interest.

Time. In WSNs usually the term *time* is associated with the network lifetime, which has a high impact on the required degree of communication and energy efficiency. However the term time can also characterize the way how the distributed processing is done. If the network is used to monitor some considered area, the application can perform *periodic* tasks to gather sensor readings. This solution is maybe not energy efficient, but collected data may be used in further analysis. Another way to monitor the same area is *event-triggered* solution – the application is characterized by two phases:

- during event detection, the system is largely quiescent, with each node monitoring the values it samples from the environment with little or no communication involved;
- if and when the event condition is met (e.g., a sensor value raises above a threshold), the WSN begins its distributed processing [6].

Obviously the provided classification is not complete. It is certainly debatable which issues are important enough to be explicitly listed and one could argue in favor of adding more dimensions. In order to categorize the various strategies for nodes positioning it is worth considering to add two more properties: *heterogeneity* and *network topology*.

Heterogeneity. Early sensor network visions anticipated that sensor networks would typically consist of homogeneous devices that were mostly identical from a hardware and software point of view [2]. However, in many applications available today, sensor networks consist of a variety of different devices. Nodes may differ in the type and number of attached sensors; some nodes may act as gateways to long-range data communication networks (e.g., GSM networks or satellite networks). The differences between nodes can be also connected with roles the nodes play in the network (some nodes my work as a cluster heads). The roles assignment can be temporary or permanent.

Network topology. Another important property of a sensor network is the maximum number of hops between any two nodes in the network. In its simplest form, a sensor network forms a single-hop network, with every sensor node being able to directly communicate with every other node or the base-station at least. In multi-hop networks nodes may forward messages over multiple hops.

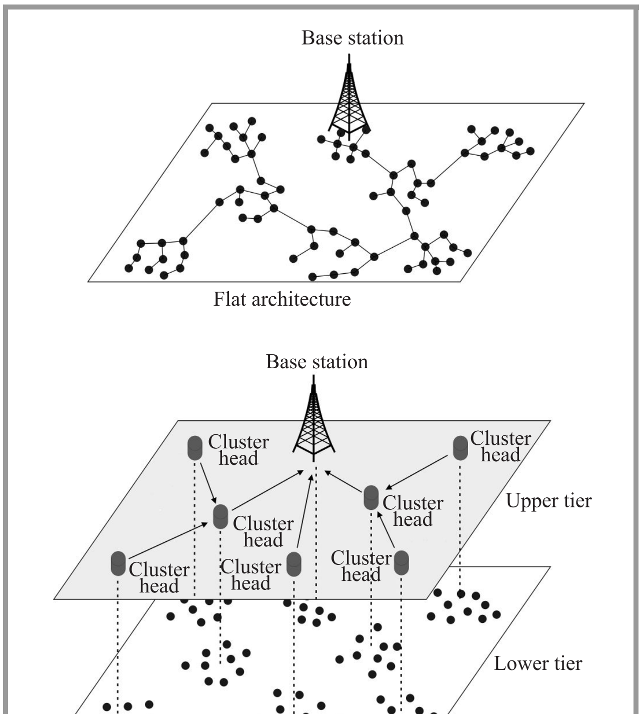


Fig. 4. Flat and tiered network topologies.

The sensor network architecture can be flat where all sensors play the same role in communication – all nodes acts as routers or it can be tired. The most common is two-tier where sensors are split into clusters; each is led by an cluster head node, as illustrated in Fig. 4.

3. Objectives

The positioning of nodes in a sensor network has received a notable attention in research. The localization and deployment are the fundamental issues and the number of papers concerning these problems exceeds few hundred. Good overview of various strategies for node placement has been provided by Younis and Akkaya [7]. They have distinguished four primary objectives for sensor deployment, such as: area coverage, network connectivity, network longevity and data fidelity. In this section we extend the list and provide a short overview of optimization objectives.

3.1. Coverage

The coverage problem is the objective that has been widely discussed in the literature. Typically considered problems are area coverage, point/target coverage, energy-efficient coverage and *k*-coverage problem. Assessing the coverage varies based on the underlying model of each sensor's field of view and the metric used to measure the collective coverage of deployed sensors.

The most commonly used sensor coverage model is a sensing disk model. All points within a disk centered at sensor are considered to be covered by the sensor. In the literature of WSNs, however, many papers assume a fixed sensing range and an isotropic detection capability of sensor. The detection ability within coverage of a sensor can be classified as the 0/1 coverage model (binary model), the probabilistic coverage model, and the information coverage model. Some of the published papers, especially early ones, use the ratio of the covered area to the size of the overall deployment region as a metric for the quality of coverage. Since 2001, however, most work has focused on the worst case coverage, usually referred to as least exposure, measuring the probability that a target would travel across an area or an event would happen without being detected [7].

3.2. Differentiated Detection Levels

Differentiated sensor network deployment, which considers the satisfaction of detection levels in different geographical characteristics, is also an important issue. In many realworld WSN applications, such as underwater sensor deployment or surveillance applications, the supervised area may require extremely high detection probabilities at certain sensitive areas. However, for some not so sensitive areas, relatively low detection probabilities are required to reduce the number of sensors deployed so as to decrease the cost. In this case, different areas require different densities of deployed nodes Therefore, the sensing requirements are not uniformly distributed within the area. As a result, the deployment strategy of WSN should take into consideration the geographical characteristics of the monitored events [8].

3.3. Network Connectivity

Another issue in WSN design is the connectivity of the network. We say that the network is connected if any active node can communicate with any other active node (possibly using other nodes as relays). Network connectivity is necessary to ensure that messages are propagated to the appropriate base station and the loss of connectivity if often treated as the end of network life. This property is strongly connected with coverage and energy efficiency (the value of transmission range may vary according to transmission power). The relationship between coverage and connectivity results from sensing and transmission ranges. If the transmission range of a node is much longer than its sensing range then connectivity is not an issue, because the coverage ensures there is a way to communicate. Situation is different if the communication range is less than sensing range.

3.4. Network Lifetime

One of the major challenges in the design of WSNs is the fact that energy resources are very limited. Recharging or replacing the battery of the sensors in the network may be difficult or impossible, causing severe limitations in the communication and processing time between all sensors in the network. Note that failure of regular sensors may not harm the overall functioning of a WSN, since neighboring sensors can take over, provided that their density is high. Therefore, the key parameter to optimize for is network lifetime – the time until the network gets partitioned in a way that is is impossible to collect the data from a part of the network [9].

3.5. Data Fidelity

Ensuring the credibility of the gathered data is obviously an important design goal of WSNs. A sensor network basically provides a collective assessment of the detected phenomena by fusing the readings of multiple independent (and sometimes heterogeneous) sensors. Data fusion boosts the fidelity of the reported incidents by lowering the probability of false alarms and of missing a detectable object. Increasing the number of sensors reporting in a particular region will surely boost the accuracy of the fused data. However, redundancy in coverage would require an increased node density, which can be undesirable due to increased *cost* or decreased *survivability* (the potential of detecting the sensors in a combat field) [7].

3.6. Energy Efficiency

This criteria is often used interchangeably with *lifetime*. Due to the limited energy resource in each sensor node, we need to utilize the sensors in an efficient manner so as to increase the lifetime of the network. There are at least two approaches to the problem of conserving energy in sensor networks connected with optimal placement. The first approach is to plan a schedule of active sensors that enables other sensors to go into a sleep mode utilizing overlaps among sensing ranges. The second approach is adjusting the sensing range of sensors for energy conservation.

3.7. Number of Nodes

This criteria is obvious. The more sensors are used the higher is cost. At least the half of the papers dedicated to optimal node deployment consider to achieve the specified goals with minimum cost.

3.8. Fault Tolerance and Load Balancing

Fault tolerant design is required to prevent individual failures from shortening network lifetime. Many authors focus on forming k-connected WSNs. K-connectivity implies that there are k independent paths among every pair of nodes. For $k >= 2$, the network can tolerate some node and link failures. Due to many-to-one interaction pattern k-connectivity is especially important design factor in the neighborhood of base stations and guarantee certain communication capacity among nodes.

4. Multi-Objective Approaches

The criteria presented in previous section are conflicting objectives (e.g., coverage versus energy consumption, fault tolerance vs survivability). Thus, there is no single nodes placement that can optimize all objectives simultaneously and a decision maker needs an optimal trade-off. In this section we provide an overview of published work according to multi-objective methods for nodes placement in wireless sensor networks. Table 1 consist the list of papers with considered objectives.

All the works except paper [10] treat a coverage as one of the objectives. Molina *et al.* have also considered coverage but as a constraint. Their aim is to obtain a full coverage network with minimum cost and maximum lifetime. The lifetime is defined as the time until the first node fails (time to first failure – TTFF). The terrain is modeled as a discrete grid, where each point in the grid represents one square meter of the terrain. They assume a nonfixed amount of homogeneous sensor nodes has to be placed in the terrain. The number of sensor nodes and their locations have to be chosen in a way that minimizes the energy spent in communications by the most loaded node in the network and the cost of the network which, in this case, is calculated as the number of deployed sensor nodes. All presented algorithm was able to find a front of non-dominated feasible solutions. Authors do not provide any model of preferences and do not select the preferred solution.

Another paper with discretized space – this time in 3D – have been written by Kang and Chen. In paper [11] N sensors are deployed to cover the sensor field. The sensor field consist of $k \times k \times k$ grid points in the x, y, z dimensions. Each sensor has an initial sensor energy and has the capability to adjust its sensor range. Sensing range options correspond to energy consumptions and detection error ranges. Three objectives are considered: maximization of coverage,

Table 1
A comparison between the various approaches for multi-objective nodes placement

No.	Title	Objective 1	Objective 2	Objective 3
1	An evolutionary approach for multi-objective 3D differentiated sensor network deployment [11]	Coverage	Differentiated detection levels	Energy efficiency
2	Layout optimization for a wireless sensor network using a multi-objective genetic algorithm [12]	Coverage	Lifetime	
3	Adaptive design optimization of wireless sensor networks using genetic algorithms [9]	Energy-related parameters	Sensing points' uniformity	
4	Multi-objective optimization for coverage control in wireless sensor network with adjustable sensing radius [13]	Coverage	Number of sensors	Energy efficiency
5	Energy efficient coverage control in wireless sensor networks based on multi-objective genetic algorithm [14]	Coverage	Number of sensors	
6	Optimal sensor network layout using multi-objective metaheuristics [10]	Energy efficiency	Number of sensors	
7	A multi-objective evolutionary algorithm for the deployment and power assignment problem in wireless sensor networks [15]	Coverage	Lifetime	
8	Multi-objective genetic algorithm for the automated planning of a wireless sensor network to monitor a critical facility [16]	Coverage	Survivability	Number of sensors

Table 2
A properties of the various approaches for multi-objective nodes placement

No.	Number of sensors	Initial deployment	Time	Heterogeneity	Network topology
1	Constant	Controlled	Event-triggered	Homogeneous	Flat
2	Constant	Controlled	Periodic	Homogeneous	Flat
3	Constant	Controlled	Periodic	Homogeneous	Two-tier
4	Variable	Existing	Periodic	Heterogeneous	Two-tier
5	Variable	Existing	Periodic	Heterogeneous	Two-tier
6	Variable	Controlled	Periodic	Homogeneous	Flat
7	Constant	Controlled	Periodic	Homogeneous	Flat
8	Variable	Controlled	Event-triggered	Homogeneous	Flat

maximization of differentiated detection levels and minimization of energy consumption. Decision variables are the 3D coordinates and the sensing ranges of all the nodes. As a final result authors present the box plots of obtained non-dominated solutions and the maximum and minimum objective values calculated in different objective functions. In paper there is no preference modeling.

Ferentinos *et al.* [9] have studied node positioning in a two-tiered network model. They concentrate on fulfilling some application specific objectives (from the scope of precision agriculture). The optimization problem is defined by the minimization of the energy-related parameters (operational energy, communication energy and battery capacity penalty) and the maximization of sensing points' uniformity, subject to the connectivity constraints and the spatial density requirement. The authors consider a cluster-based network architecture and a constant number on nodes. Unfortunately the provided solution cannot be treated as a multi-objective one, because all objectives was combined into single objective function (weighted sum approach).

Two-tiered architecture have been also considered by Jia *et al.* in papers [14] and [13]. The former paper is dedicated to optimal coverage control scheme in existing network. There are two objectives: maximization of coverage and minimization of number of sensors. In the paper [13], the problem of maintaining sensing coverage by keeping a small number of active sensor nodes and a small amount of energy consumption is studied. This time the list of objectives has been extended to include energy efficiency. Both papers show an interesting studies of multi-objective optimization. However Jia *et al.* consider slightly different task, because they optimize an existing network, so the nodes placement cannot be treated as decision variable. More information about differences in network properties for considered papers can be found in Table 2.

Typical trade-off between area coverage and network lifetime has been considered in papers [12], [15]. In both papers the considered area is a flat square surface where sensor nodes can monitor anything within R_{sensor}, and where they can communicate with any other node located within R_{comm}. In paper [12] the base station, with which every sensor must communicate (either directly or via hops through nearby sensors), is placed in the center of the area. Each sensor initially has the same energy available in its battery, and it is assumed that energy decreases by one arbitrary unit for every data transmission. The design variables are the 2D coordinates of the sensors. In paper [15] there is one additional vector of decision variables connected with the transmission power level of each sensor. Two objectives are considered: maximization of coverage and maximization of lifetime. As a final result authors present a Pareto front from which the user can choose.

Similar assumption about network configuration has been assumed in paper [16]. The sensors are identical and are placed in a flat square. The design variables are the 2D coordinates of the sensors. Three examples has been described by Jourdan and de Weck. The most interesting is monitoring movements in and aut of a facility served by two roads. The first objective is the coverage, by which is meant the ability of the network to monitor movements in and out of the facility. The second objective is the survivability of the network, by which is meant the likelihood that sensors will not be found. Each point in the area is assigned a probability of detection. This probability depends on the proximity of the facility or the roads. It is assumed that if a sensor is placed close to a road (where most of the activity takes place) or to the facility, it is more likely to be found and disabled. The third objective is the number of sensors. As a final result authors present a set of non-dominated solutions.

5. Summary and Conclusions

In this paper we outline the main properties and criteria that should be considered while deploying the nodes in considered area. We provided an overview of multi-objective strategies, their assumptions, optimization problem formulation and results. All the authors concentrate on optimization methods and finding a Pareto frontier. The model of preferences was not present in any paper and authors did not try to select the preferred solution. More work is required in order to provide the solution which can be applied in real applications.

Acknowledgements

This work was supported by Ministry of Science and Higher Education grant NN514 416934.

References

[1] C. G. Cassandras and W. Li, "Sensor networks and cooperative control", *Eur. J. Control*, vol. 11, no. 4–5, pp. 436–463, 2005.

[2] K. Romer and F. Mattern, "The design space of wireless sensor networks", *IEEE Wirel. Commun.*, vol. 11, no. 6, pp. 54–61, 2004.

[3] Y. Wang, H. Dang, and H. Wu, "A survey on analytic studies of delay-tolerant mobile sensor networks: research articles", *Wirel. Commun. Mob. Comput.*, vol. 7, no. 10, pp. 1197–1208, 2007.

[4] I. F. Akyildiz and I. H. Kasimoglu, "Wireless sensor and actor networks: research challenges", *Ad Hoc Netw.*, vol. 2, no. 4, pp. 351–367, 2004.

[5] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks", *Comput. Netw.*, vol. 51, no. 4, pp. 921–960, 2007.

[6] L. Mottola and G. P. Picco, "Programming wireless sensor networks: fundamental concepts and state-of-the-art", *ACM Comput. Surv.*, 2009 (to appear).

[7] M. Younis and K. Akkaya, "Strategies and techniques for node placement in wireless sensor networks: a survey", *Ad Hoc Netw.*, vol. 6, no. 4, pp. 621–655, 2008.

[8] J. Zhang, T. Yan, and S.H. Son, "Deployment strategies for differentiated detection in wireless sensor networks", in *Proc. IEEE SECON 2006*, Reston, USA, 2006, vol. 1, pp. 316–325, 2006.

[9] K. P. Ferentinos and T. A. Tsiligiridis, "Adaptive design optimization of wireless sensor networks using genetic algorithms", *Comput. Netw.*, vol. 51, no. 4, pp. 1031–1051, 2007.

[10] G. Molina, E. Alba, and E.-G. Talbi, "Optimal sensor network layout using multi-objective metaheuristics", *J. Univer. Comput. Sci.*, vol. 14, no. 15, pp. 2549–2565, 2008.

[11] C. Kang and J. Chen, "An evolutionary approach for multi-objective 3d differentiated sensor network deployment", in *Proc. IEEE Int. Conf. CSE'09*, Vancouver, Canada, 2009, vol. 1, pp. 187–193.

[12] D. B. Jourdan and O. L. de Weck, "Layout optimization for a wireless sensor network using a multi-objective genetic algorithm", in *IEEE Veh. Technol. Conf. P.*, vol. 5, pp. 2466–2470, 2004.

[13] J. Jia, J. Chen, G. Chang, Y. Wen, and J. Song, "Multi-objective optimization for coverage control in wireless sensor network with adjustable sensing radius", *Comput. Math. Appl.*, vol. 57, no. 11–12, pp. 1767–1775, 2009.

[14] J. Jia, J. Chen, G. Chang, and Z. Tan, "Energy efficient coverage control in wireless sensor networks based on multi-objective genetic algorithm", *Comput. Math. Appl.*, vol. 57, no. 11–12, pp. 1756–1766, 2009.

[15] A. Konstantinidis, K. Yang, Q. Zhang, and D. Zeinalipour-Yazti, "A multi-objective evolutionary algorithm for the deployment and power assignment problem in wireless sensor networks", *Comput. Netw.*, 2009 (in press).

[16] D. B. Jourdan and O. L. de Weck, "Multi-objective genetic algorithm for the automated planning of a wireless sensor network to monitor a critical facility", in *SPIE Proc.*, vol. 5403, pp. 565–575, 2004.

Michał Marks received his M.Sc. in computer science from the Warsaw University of Technology, Poland, in 2007. Currently he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2007 he works at the Research and Academic Computer Network (NASK). His research area focuses on global optimization, multiple criteria optimization, decision support and machine learning.

e-mail: mmarks@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

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

Paper

Personal Ontologies for Knowledge Acquisition and Sharing in Collaborative PrOnto Framework

Cezary Chudzian and Jarosław Sobieszek

Abstract—This paper summarizes our preliminary experiences with implementing some of the ideas lying behind the concept of creative environment. Research group at the National Institute of Telecommunications has developed a prototype framework for collaborative knowledge acquisition and sharing, called PrOnto. At the moment the artifacts we organize and share are typical sources of scientific knowledge, namely journal papers and web pages. In PrOnto we introduce two interrelated explicit levels of knowledge representation: keywords and ontological concepts. Each user of the framework maintains his own ontological profile, consisting of concepts and each concept is, in turn, by subjective user's decision, related to a set of weighted keywords that define its meaning. Furthermore, dedicated indexing engine is responsible for objectively establishing correspondence between documents and keywords, or in other words, the measure of representativeness of the keyword to document's content. Developing an appropriate knowledge model is a preliminary step to share it efficiently. We believe that higher level representation facilitates exploration of other people's areas of interest. PrOnto gives an opportunity to browse knowledge artifacts from the conceptual point of view of any user registered in the system. The paper presents the ideas behind the PrOnto framework, gives an outline of its components and finalizes with a number of conclusions and proposals for future enhancements.

Keywords*—*collaborative knowledge sharing, creativity support, knowledge acquisition, knowledge management, ontologies.

1. Introduction

Willing to support the development of knowledge creating environments, one has to consider common patterns existing in knowledge intensive processes maintained at, not only academic and research institutions, but also growing number of commercial companies, trying to improve their position in contemporary knowledge based economy market, by putting higher stress on knowledge management activities. Instantiations of those patterns differ from institution to institution, depending on the maturity of knowledge management policy development, but still they can be observed at, at least, rudimentary stage.

Models of creative processes have been investigated for many years now. Significant milestone on the pathway of research in this area has been put by Nonaka and Takeuchi in [1]. They introduced SECI[1] spiral as an algorithmic model of organizational knowledge creation. Theory of Nonaka and Takeuchi describes the creation process as reliant on the collaboration of individuals involved and shows the special role of knowledge transfers between implicit, codified representation and tacit, intuitive form. Concepts of Nonaka and Takeuchi have been widespread in and met interest of the knowledge management community. Here, in this paper, we refer to the further augmented theory of creative environment, better suited to creative environments, namely triple helix of normal knowledge creation presented by Wierzbicki and Nakamori in [2]. Triple helix is the combination of three spirals modeling three aspects of knowledge creation: hermeneutic, experimental and intersubjective. All they have a cross-cutting point – enlightenment – a transition in creative space, expressing the creation of the new idea.

The enlightenment-analysis-hermeneutical immersion-reflection (EAIR) spiral, reflects the process of searching through rational heritage of humanity and reflecting on the object of study. It is usually accomplished with a repetitive "search & browse" strategy, usually implemented in the way as follows. First some query against knowledge repository is performed and after browsing over the results, selection of relevant information and drawing new conclusions, refined query is prepared that starts another strategy iteration. More specific form of the strategy is acquiring new knowledge through reading scientific papers. Starting with a very rough idea on the object of study, one looks up for papers with the keywords and titles somehow corresponding with the object. The more papers one reads, the more accurate his query may be and, in turn, more appropriate knowledge resources one can find in subsequent steps of "search & browse" run. Unfortunately, there exists a risk of obtaining the query overfit to what one already knows, making it harder to find any new, but relevant resources. Wider exploration of the research area may be facilitated by looking at the object of study from different individual perspectives, thus extending the "search & browse" paradigm with some collaborative dimension. We shall consider this important issue further on.

The enlightenment-experiment-interpretation-selection (EEIS) spiral models verification and objectification of the ideas through scientific experiments. As we do not provide support for this area of knowledge creation in our developed framework, we only mention here its existence, without going into details, which may be found in [2].

[1] Acronym for the names of transitions present in the model (socialization externalization combination internalization).

Debating on the ideas obtained from other spirals or through any other source of enlightenment is a subject of the enlightenment-debate-immersion-selection (EDIS) spiral. Implementing process modeled with EDIS is crucial for tacit knowledge sharing and encompasses transitions between tacit and explicit levels of knowledge and between group and individuals. The framework we present herein supports EDIS spiral in indirect way. First, it helps to meet people interested in common topics, second it facilitates acquiring and sharing textual materials for the debate.

In the following sections we present work of other teams done in our area of interest, then we provide more formal definition of the knowledge representation we use, present knowledge sharing capabilities of the framework and finally conclude with the steps to be made in the future.

2. Related Work

The problems of knowledge acquisition, organizing and sharing have recently gained much attention. Systematic review of the developed solutions and already finished or still running projects is far out of scope of this paper. Nevertheless, we will try to provide the reader with information on some selected tools and methods, we have examined throughout our research. We were especially interested in software products that organize knowledge around structures more complex than just bag of keywords and leverage cooperation between individuals for effective knowledge acquisition and sharing.

The most common way of performing "search & browse" routine, as mentioned above, is supported by one of general purpose or dedicated search engines and usually is organized as repetitive query refinement on the basis of previous findings. Query is, in fact, a set of keywords. Leading companies on the search market have already noticed that keyword search is getting less effective with the growth of available information amount and new approaches to finding and structuring information are needed. Therefore they started to work on the new products, closer to the idea of semantic search. In June, 2009 Google launched an experimental service, called Squared, which displays search results in a tabular form, with rows representing objects and columns corresponding to their common attributes. One month earlier, Stephen Wolfram[2] released his Wolfram Alpha answering engine, with queries interpreted semantically, before giving the answers drawn from underlying, structured knowledge base.

Growing popularity of social network services creates a new potential for structuring and personalizing knowledge resources. The biggest service of this kind, Facebook, with 200 million users storing their data on Facebook servers, may be perceived as an alternative web [3]. Its power comes from the fact that, in contradiction to the web, it keeps its content organized and personalized from the very beginning, when the piece of information is shared by the user. Much less expanded social networks, like Index Copernicus, BiomedExperts or BioCrowd, have been developed to facilitate knowledge sharing and organizing communities of practice focused on common topics.

The vast majority of web search engines, as well as social networks, assess relevance of a piece of information to the object of interest, on the basis of some keyword-based model. In general there are two basic approaches. One is to define some objective measure of relevance, for instance, the number of occurrences of every keyword found in the text document[3] and rank documents according to its value. On the other pole one finds a subjective model in which person annotates pieces of information with keywords of one's choice - so called tags. Both those models, in classical form, do not organize keywords in any semantic structure, using them as ordinary textual labels.

Combining richer indexing models, specifically ontology-based ones, with social networking, in order to develop novel knowledge management tools has been a subject of investigation in research projects for a couple of years now. Social networking contributes its value – further dimension of the knowledge space – as every piece of information is associated with its contributor. Ontologies, defined as a "*formal, explicit specification of a shared conceptualization*", create semantic backbone, linking resources of parties involved and organizing them around common conceptual structures.

OntoShare [4], a tool for knowledge sharing within communities of practice, is one of the examples. Common ontology of the group is agreed upon and imported into the system. Each community member contributes textual documents he judges as relevant to the interests of the whole group. The semantic proximity between the concepts from ontology and documents is measured on the basis of their profiles. Document's profile and ontological concept's profile are sets of keywords with weights measuring how much given keyword is representative to corresponding document or concept. The weights and keywords are computed by a specialized background algorithm and they are not explicitly exposed to the user. OntoShare user subscribes to existing concepts, thus adding them to his own profile and tags documents with concepts' signatures. The latter indirectly influences the profile of the concept as it is the main input of the computing algorithm. The OntoShare way of building ontological structure is called usage-based evolution of the ontology. The primary usage scenarios are document recommendation and finding users with similar interests to facilitate tacit knowledge sharing. They are both accomplished with the use of the acquired profiles.

PrOnto shares some of the ideas implemented in OntoShare. There are, however, important differences between them in the way the ontology is defined and maintained and how they deal with the keywords and relate them to ontological profiles, not to mention disparate interfaces for human – computer interaction. Moreover OntoShare is no

[2] Known previously as mathematica's author.

[3] It is called term frequency and is well known in the community of text miners.

longer available in the public domain, at least it is not accessible from project dedicated website.

The SWAP[4] project [5], [6] is another example of EU-financed project situated in the area of knowledge management through application of ontological models in networked environments. The network is decentralized in a peer-to-peer manner, which promises greater scalability. The semantic concepts are specific to every node (user) of the network and ontology matching techniques are applied to discover the grade of correspondence. The only known and available instance of SWAP-like system is Bibster, peer-to-peer network for sharing bibliographical information [7].

Recapitulating, there is a lack of publicly available knowledge management tools, organizing knowledge artifacts around structures more expressive and human understandable than simple keywords, facilitating knowledge sharing and leveraging the power of social networking. Therefore to address those issues we have decided to develop PrOnto.

3. Motivation

The main goal we were aiming at was to create a social networking platform for organizing and sharing knowledge resources by leveraging activities of network members to collect and index resources and to accelerate "search & browse" processes, thus supporting hermeneutic EAIR spiral execution. We wished to build up a digital library of documents with a certain level of quality assured. Collected artifacts ought to be accessible by every single user of the platform from his own semantic perspective. Further, formal representation of the perspective maintained by the user should be available to his colleagues as well, in order to facilitate cooperation and to speed up their learning processes in the areas they do not know, but which had been already investigated by their colleagues. The knowledge structure was to be organized in the way that not only let people order existing library items, but also was capable of accumulating new knowledge, fitting new documents to the structure, thus making it possible for the user to discover previously unknown, but relevant resources.

We have started our work on PrOnto framework having in mind some general rules and remarks, coming from previous experience, intuition and common sense. We have been following them then as the development guidelines. Let us discuss them shortly as they have influenced the current shape of the framework.

First observation is that semantically richer indexing schemes, specifically ontologies, enable contextual access to knowledge resources and thus allow their more intuitive exploration and, in turn, support cognitive processes. Still appropriate presentation layer has to be proposed, leveraging ontology-based knowledge representation. Particularly suited for interactive systems, such as PrOnto, is the visual form presentation. Diagramming approaches, like semantic networks [8], mind mapping [9], concept mapping [10], have proven their usefulness in human-oriented modeling of conceptual areas. They facilitate understanding and accelerate learning processes.

It seems reasonable to think that every human being feels more comfortable arranging his knowledge according to his own conceptual structure. Personalized ontological perspective may then serve as a guide to a subdomain of knowledge, recognized and arranged by a person, for other people use, especially when it is presented in a handy visual form. On the other hand, using ontologies as the knowledge representation means to have a common conceptualization of the domain. Therefore, while maintaining individual ontologies, it is essential to provide users with a set of tools facilitating ontology matching.

Semantically overlapping content can be usually retrieved with many different keyword queries. An example may be the concept of uplift modeling, being a predictive modeling technique. According to the information provided by Nicholas Radcliffe [11], one of its inventors, more than eight keyword queries characterize information on the concept. Those are: uplift modeling, differential response analysis, incremental modeling, incremental impact modeling, true response modeling, true lift modeling, proportional hazards modeling, net modeling. Using every one of them as a query in any web search engine results in different set of web pages retrieved, but the content is semantically close. Someone who is not familiar with that domain, which is typical case when he is just about to start exploring it, will have less chance to get relevant and useful information. Sharing queries, not only the artifacts itselves, can therefore support much wider exploration.

The high quality of information is an important factor for the knowledge creating environment. Creating a digital library out of knowledge sources recommended by, to some extent, trusted person might turn the social network into a filtering engine for quality control. Every piece of information becomes a part of the library by a conscious decision of the recommender.

4. Knowledge Representation

Before going into details of knowledge representation model we implemented in PrOnto framework, some attention has to be paid to a concept of hermeneutic horizon. In PrOnto, and further in this paper, we use the word "horizon" when referring to individual ontological profile, being a personalized perspective imposed on some domain of interest. Any user or a group may organize knowledge around their own semantic structure, or in PrOnto terminology, horizon.

But the term hermeneutic horizon has even deeper philosophical implications. Gadamer [12] defined it as: *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*.

Alternative definition by modern Polish philosopher Król [13], says *the hermeneutic horizon is a set of intuitive*

[4] Acronym for *semantic web and peer-to-peer*.

assumptions on the object of study. PrOnto's way of understanding the horizon is closer to the meaning developed by Gadamer, as it refers more to explicit level of knowledge, instead of implicit, intuitive one.

Schema of the knowledge structure implemented in PrOnto framework is illustrated in Fig. 1. It consists of three levels of representation: artifacts (documents) D – keywords K – horizons (profiles) H.

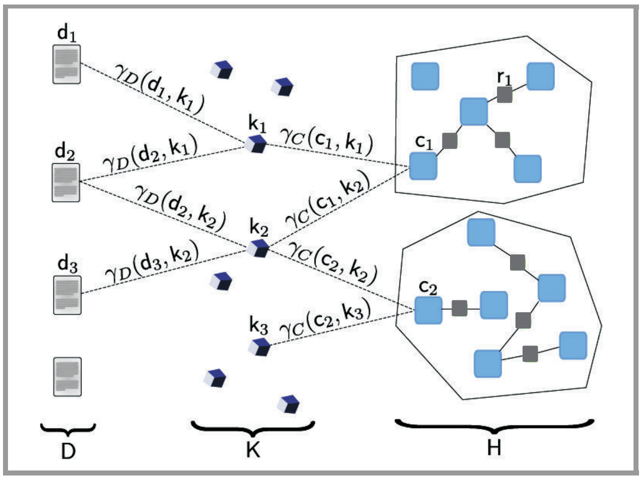


Fig. 1. Knowledge structure in PrOnto.

Definition 1: Knowledge in PrOnto framework is organized around the structure:

$$\mathsf{KR} := (\mathsf{H},\mathsf{C},\mathsf{R},\mathsf{K},\mathsf{D},\alpha_\mathsf{C},\alpha_\mathsf{R},\sigma,\gamma_\mathsf{C},\gamma_\mathsf{D})\,, \qquad (1)$$

where:

- C is a set of concepts uniquely identified within the framework. In contradiction concepts' names or labels are not required to be unique.
- R is a set of relations. Every relation is unique, but the labels of the relations might repeat.
- $\sigma : \mathsf{R} \mapsto \mathsf{C} \times \mathsf{C}$ is a mapping that specifies concepts for which the relation holds.
- H is a set of horizons. Horizon is an individual perspective superimposed on the knowledge accumulated in the system. Every concept and relation is localized within a single horizon, which is reflected by the following mappings:
  - $\alpha_\mathsf{C} : \mathsf{C} \mapsto \mathsf{H}$
  - $\alpha_\mathsf{R} : \mathsf{R} \mapsto \mathsf{H}$
- K is a set of keywords. Keyword is an ordered set of words in a fixed grammatical form.
- D is a set of knowledge artifacts. Currently PrOnto framework deals only with textual documents, thus further we will be using term document interchangeably.
- $\gamma_\mathsf{D} : \mathsf{D} \times \mathsf{K} \mapsto \Re$ is a function measuring how strongly a keyword $\mathsf{k} \in \mathsf{K}$ represents an artifact $\mathsf{d} \in \mathsf{D}$, given fixed collection D.
- $\gamma_\mathsf{C} : \mathsf{C} \times \mathsf{K} \mapsto \Re$ is a function, measuring how much a keyword contributes to the meaning of a concept according to the preferences defined for the horizon $\alpha_\mathsf{C}(\mathsf{c})$ within which the concept has been defined. The measure corresponds to the conditional probability $P(c|k)$, $\mathsf{c} \in \mathsf{C}$ and $\mathsf{k} \in \mathsf{K}$.

4.1. Implemented Measures

Although the basic model does not make any assumptions on the formal, mathematical definitions of γ_D and γ_C, we had to decide on some specific implementation for the purpose of PrOnto development.

γ_D is to be an objective measure, reflecting both, strength of relation between k and artifact d and how d is distinguished among other artifacts with respect to k, or in other words how k is representative to d and unrepresentative to $\mathsf{D}\setminus\{\mathsf{d}\}$. As the current version of PrOnto limits artifacts to textual documents, we have adopted TF-IDF measure as γ_D function. TF-IDF stands for term frequency – inverse document frequency and is well-known tool in the text mining and information retrieval community for measuring document's relevance to a given query.

$$\text{TF-IDF}(\mathsf{k},\mathsf{d}) = \frac{\mu(\mathsf{k},\mathsf{d})}{|\{\mathsf{k}':\mathsf{k}'\in\mathsf{K}\wedge\mathsf{k}'\in_\mathsf{k}\mathsf{d}\}|}\cdot\log\frac{|\mathsf{D}|}{|\{\mathsf{d}':\mathsf{k}\in_\mathsf{k}\mathsf{d}'\}|}\,, \qquad (2)$$

with $\mu(\mathsf{k}',\mathsf{d}')$ being a number of occurrences of k' in d'. Relation $\in_\mathsf{k}$ denotes "k occurs in d". See [14] for more information on term weighting approaches in information retrieval.

Relation between ontological concept and keyword is, on the other hand, measured subjectively. The user is equipped with an interactive tool for adjusting the strength of every concept-keyword relation by picking a value from some predefined interval. While PrOnto approach is completely manual and thus subjective, alternative procedures have been also proposed, like those implemented in OntoShare or OntoGen [15] systems. They derive concept profiles as keyword vectors, by analyzing document corpus in a semi-automatic fashion. We consider adding such a procedure as a further extension to our prototype framework, but still leaving the final decision to the user.

4.2. Ranking Method

Given two above measures, γ_C and γ_D, one can construct ranking procedure, for ordering knowledge artifacts from D according to their relevance to the concept $\mathsf{c} \in \mathsf{C}$. Obviously, any artifact is tied to a concept through a set of common keywords and there are many ways to leverage this indirect association for ranking definition. In the current stage PrOnto ranks documents in concept perspective, utilizing easy to compute in a database, and conceptually simple function ϕ.

Definition 2: Ranking function $\phi : \mathsf{D} \times \mathsf{C} \mapsto \Re$ takes form:

$$\phi(\mathsf{c},\mathsf{d}) = \sum_{\mathsf{k} \in \mathsf{K}} \gamma_{\mathsf{C}}(\mathsf{c},\mathsf{k}) \cdot \gamma_{\mathsf{D}}(\mathsf{d},\mathsf{k}), \ \forall \mathsf{c} \in \mathsf{C}, \ \forall \mathsf{d} \in \mathsf{D}. \quad (3)$$

Interpretation is rather straightforward. We shall only notice the number of ranking procedures one can adapt here is much bigger, ranging from simple counting of common keywords to complex interactive multicriteria analysis.

5. Knowledge Sharing

PrOnto is based on a client-server architecture with a client-side application running inside a web browser and central server storing all the metadata and the library of collected knowledge resources. Upload of documents is implemented as a firefox browser extension. Client application, developed using flash technology, allows editing concept maps, adding new keywords and linking them with the concepts, searching and browsing the library, receiving alert messages on significant events occurring in the system.

In this section we present in more details how knowledge sharing is realized within PrOnto framework. Our discussion is organized around three main subtopics, corresponding to different levels of knowledge representation. First is exchanging artifacts, second is sharing procedures of locating them and third is about finding someone who is likely to know that procedure.

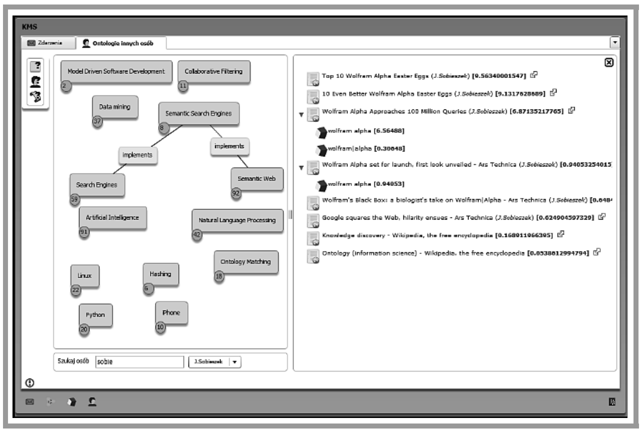
5.1. Sharing Artifacts

While searching and browsing the web, user may take a conscious decision to share a piece of information with other framework users. Firefox browser extension is used as an entry point for document delivery. At the time decision is being made, the document becomes a part of the repository and then dedicated module takes care of extracting the most relevant keywords, computing γ_{D} measures. From then on it is accessible for any user, by any access method implemented within the framework.

The main view on the conceptual horizon (Fig. 2) is implemented as a concept map-like graph, with concepts as nodes connected with named relations. Given ranking procedure realized with scalarizing function ϕ Eq. (3), there exists ordering of documents for every concept. User's own graph is fully editable, others are accessible in the read-only mode, letting the user to browse knowledge resources from any semantic perspective defined within the framework.

The access to the library through multidimensional "search & browse" view (Fig. 3) is closer to standard search engine approaches, however it enables additional semantic features to be added to search criteria or as browsing dimensions. Exploiting direct or indirect relations between knowledge model components one can analyze, for instance, which concepts are covered by the document content and what keywords are common to both concept and document (see Fig. 2).

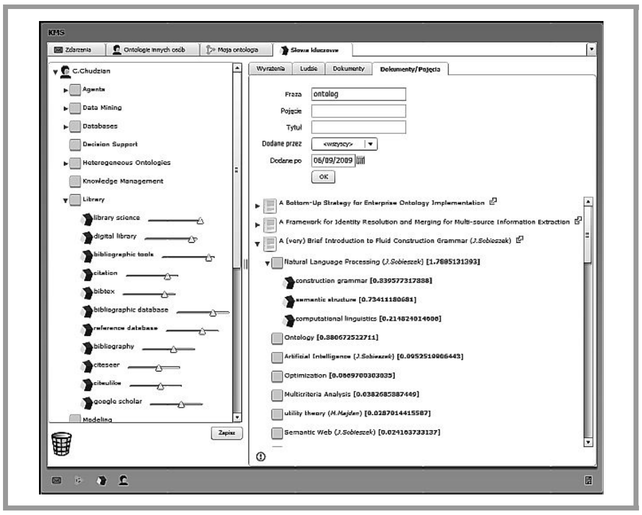
In PrOnto, there exists a mechanism, acting like a subscription service. Each time a new knowledge artifact is added



Szukaj osób sobie J.Sobieszek

Top 10 Wolfram Alpha Easter Eggs (*J.Sobieszek*) [9.56340001547]
10 Even Better Wolfram Alpha Easter Eggs (*J.Sobieszek*) [9.13176288689]
Wolfram Alpha Approaches 100 Million Queries (*J.Sobieszek*) [6.87135217765]
wolfram alpha [6.56488]
wolfram|alpha [0.30648]
Wolfram Alpha set for launch, first look unveiled - Ars Technica (*J.Sobieszek*) [0.94053254015
wolfram alpha [0.94053]
Wolfram's Black Box: a biologist's take on Wolfram|Alpha - Ars Technica (*J.Sobieszek*) [0.648
Google squares the Web, hilarity ensues - Ars Technica (*J.Sobieszek*) [0.624904597329]
Knowledge discovery - Wikipedia, the free encyclopedia [0.168911966395]
Ontology (information science) - Wikipedia, the free encyclopedia [0.0538812994794]

Fig. 2. Concept map view.



C.Chudzian
Agents
Data Mining
Databases
Decision Support
Heterogeneous Ontologies
Knowledge Management
Library
library science
digital library
bibliographic tools
citation
bibtex
bibliographic database
reference database
bibliography
citeseer
citeulike
google scholar
Modeling

Fraza: ontolog
Pojęcie:
Tytuł:
Dodane przez: <wszyscy>
Dodane po: 06/09/2009

A Bottom-Up Strategy for Enterprise Ontology Implementation
A Framework for Identity Resolution and Merging for Multi-source Information Extraction
A (very) Brief Introduction to Fluid Construction Grammar (*J.Sobieszek*)
Natural Language Processing (*J.Sobieszek*) [1.7885131393]
construction grammar [0.839577317888]
semantic structure [0.73411180681]
computational linguistics [0.214824014606]
Ontology [0.880672522711]
Artificial Intelligence (*J.Sobieszek*) [0.0952510906443]
Optimization [0.0869700303035]
Multicriteria Analysis [0.0382685887449]
utility theory (*M.Majdan*) [0.0287014415587]
Semantic Web (*J.Sobieszek*) [0.024163733137]

Fig. 3. Multidimensional "search & browse" view.

to the library, users whose profiles contain matching concepts with $\phi > 0$, are alerted with a message sent to their private mailboxes.

5.2. Sharing Queries

As mentioned before, sharing knowledge is not only about creating a common repository of knowledge resources, but also about sharing queries, or in other words, procedures of finding the resources most wanted at the given moment. The basic building block of a query in its classical search engine meaning is a keyword. In PrOnto we keep keywords bound to ontological concepts of individual horizons (see the left pane in Fig. 2). As the horizons are exposed to all members of the PrOnto network, one can discover new keywords, while exploring higher level - ontological description of the domain.

Keywords are initially imported to the framework's database from any external source (e.g., Wikipedia) and then used for indexing documents flowing into the system. Just exactly as in the case of sharing documents, user can share a keyword that becomes a part of a common collection visible to all PrOnto users. Browsing through a concept map of another user, one can possibly discover new keywords, previously unknown or unrealized, that might be useful in formulation of more accurate queries. Another context that the new keyword might be recognized in, is browsing the artifacts in the framework's library. Keyword gets a high γ_D-score for the document it is relevant to and becomes visible on the list of document's characteristic keywords. So, with the mediation of library item, a keyword is transferred between users and query sharing mechanism is established.

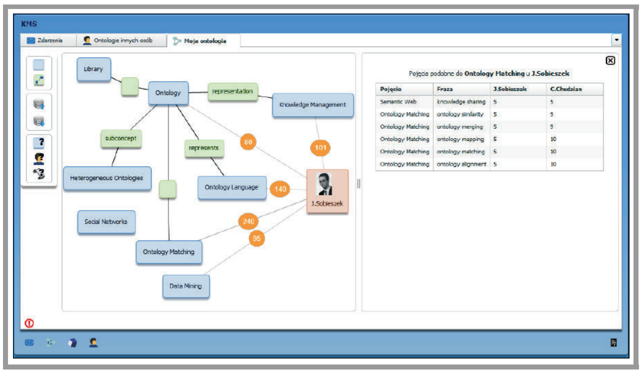
To keep the user on track of what is going on in the system, messaging module alerts the user whenever any new document is shared, or any new concept is created in the system, that is strongly related to the keyword might be interested in.

5.3. Sharing Expertise

Third, conceptually the highest level of knowledge sharing in PrOnto, is about locating domain experts for further debating on the object of study, thus supporting EDIS spiral and creating platform for tacit knowledge exchange. Since PrOnto lets every user to use his own, individual concepts, a tool must be provided for searching for the concepts semantically similar to any given one. This task has been a subject of interest within ontology matching stream of research and methods have been developed to deal with it [16]. PrOnto prototype is as far limited to assessment of the similarity between concepts by exact matching their label names and by the comparison of keywords associated with them. The latter similarity degree is measured with the formula

$$\mathrm{sim}(\mathsf{c}_i,\mathsf{c}_j) = \sum_{\mathsf{k}\in\mathsf{K}} \gamma_{\mathsf{C}}(\mathsf{c}_i,\mathsf{k})\cdot\gamma_{\mathsf{C}}(\mathsf{c}_j,\mathsf{k})\,.$$

Owner of the horizon containing concepts similar, in the sense of one of above definitions, to the ones from user's own horizon is put on the concept map view screen (Fig. 4). The multidimensional "search & browse" view marks concepts and documents with the names of their owners. Here we understand document owner as a user who shared the document uploading it with firefox browser extension.



Pojęcia podobne do **Ontology Matching** u **J.Sobieszek**

Pojęcie	Fraza	J.Sobieszek	C.Chudzian
Semantic Web	knowledge sharing	5	5
Ontology Matching	ontology similarity	5	9
Ontology Matching	ontology merging	5	9
Ontology Matching	ontology mapping	5	10
Ontology Matching	ontology matching	5	10
Ontology Matching	ontology alignment	5	10

Fig. 4. People sharing concepts.

6. Evaluation

Being the social system and applying subjective preference model for concept definition (γ_C measure), PrOnto needs an evaluation procedure adapted to those characteristics. We plan to ask users to give us a feedback on their perception of the framework. We have not yet started evaluation process. The only thing we have done in the testing area was implementation of contextual notes system. On every screen, there is a button for opening a window in which user may write down a note and categorize it with problem type and priority. The notes system covers the problems of rather technical nature. There is still a need for more formal evaluation and we plan to provide users with a questionnaire letting them to express their opinion being guided with a set of questions on usefulness and usability of the PrOnto framework.

7. Conclusions

The paper introduces PrOnto, the web based framework for acquiring and sharing knowledge artifacts. PrOnto is social networking platform whose main ambition is to support creative processes within community of practice. The knowledge in the framework is to be searched and shared at the higher, conceptual level, aiming beyond keyword based searching and sharing techniques. The user is provided with graphical interface for defining and exploring the knowledge structure.

At the moment of writing this paper PrOnto is at the prototype stage. Below we present some of the ideas for further development.

- Concept map-like structure we have implemented is a semantically weak language for describing hermeneutic horizon. The language shall be semantically strengthened for more formal description of knowledge structure.
- The γ_C measure, for subjectively associating concepts with keywords, is defined in a manual procedure. Incorporating techniques of automatic or semi-automatic estimation of initial values of γ_C on the basis of social network data and library contents would be a helpful hint tool for the users.
- We consider the query sharing task particularly interesting and important for searching the web. Employing the potential of social system for constantly improving the search procedure by making the queries more accurate adds a social dimension to the idea of hermeneutic agent [17]. We wish to explore this research direction particularly.
- The only ranking method for ordering knowledge artifacts in PrOnto is by applying scalarizing function ϕ. The model presented above, however, gives an opportunity for more complex procedures to be used, specifically interactive multicriteria analysis methods. This direction shall be examined, as well.

- The reconstruction of the learning processes performed by other users is another challenge for the future. Having learning path recorded as a sequence of steps leading to the current state of knowledge, with possibility of highlighting milestones and warning about dead ends, may accelerate knowledge acquisition.

References

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

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

[3] F. Vogelstein, "Great wall of facebook: the social network's plan to dominate the internet – and keep google out", *Wired Mag.*, iss. 17.07, 2009.

[4] J. Davies, A. Duke, and Y. Sure, "Ontoshare – an ontology-based knowledge sharing system for virtual communities of practice", *J. Univ. Comput. Sci.*, vol. 10, no. 3, pp. 262–283, 2004.

[5] M. Ehrig, P. Haase, B. Schnizler, S. Staab, C. Tempich, R. Siebes, and H. Stuckenschmidt, "Swap: semantic web and peer-to-peer project deliverable 3.6 refined methods", 2003 [Online]. Available: http://swap.semanticweb.org/public/Publications/swap-d3.6.pdf .

[6] M. Ehrig, C. Tempich, and Z. Aleksovski, "Swap: Semantic web and peer-to-peer project deliverable 4.7 final tools", 2004 [Online]. Available: http://swap.semanticweb.org/public/public/Publications/swap-d4.7.pdf

[7] J. Broekstra, M. Ehrig, P. Haase, F. Van Harmelen, M. Menken, P. Mika, B. Schnizler, and R. Siebes, "Bibster – a semantics-based bibliographic peer-to-peer system", in *Proc. 3rd Int. Seman. Web Conf.*, Hiroshima, Japan, 2004, pp. 122–136.

[8] A. Borgida and J. F. Sowa, *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo: Morgan Kaufmann, 1991.

[9] T. Buzan and B. Buzan, *The Mind Map Book*. Harlow: BBC Active, 2003.

[10] J. D. Novak, *Learning, Creating, and Using Knowledge: Concept Maps As Facilitative Tools in Schools and Corporations*. Mahvah: Lawrence Erlbaum Associates, 1998.

[11] N. J. Radcliffe, The Scientific Marketer, Uplift modeling FAQ, 2007 [Online]. Available: http://scientificmarketer.com/2007/09/uplift-modelling-faq.html

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

[13] Z. Król, *Platon i podstawy matematyki współczesnej*. Nowa Wieś: Wydawnictwo Rolewski, 2005 (in Polish).

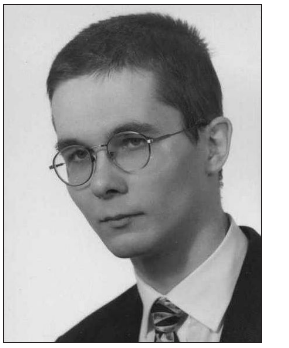
[14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Inform. Proces. Manage.*, vol. 24, iss. 5, pp. 513–523, 1988.

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

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

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

Cezary Chudzian received his M.Sc. in computer science from the Warsaw University of Technology 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



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

A Software Platform for Global Optimization

Ewa Niewiadomska-Szynkiewicz and Michał Marks

Abstract—This paper addresses issues associated with the global optimization algorithms, which are methods to find optimal solutions for given problems. It focuses on an integrated software environment – global optimization object-oriented library (GOOL), which provides the graphical user interface together with the library of solvers for convex and nonconvex, unconstrained and constrained problems. We describe the design, performance and possible applications of the GOOL system. The practical example – price management problem – is provided to illustrate the effectiveness and range of applications of our software tool.

Keywords—global optimization, integrated software systems, nonconvex optimization, numerical libraries, price management.

1. Introduction

Many decision problems are formulated as optimization tasks in which the objective function is nonconvex and has multiple extrema in the region of interest. In addition, in many practical contexts, the optimization problem cannot be described analytically due to the natural complexity and uncertainty of real-life systems. In such cases the simulation experiment is usually used to evaluate the expected performance of the system for each set of decision variables. It involves of simulation-based optimization that is the merging of optimization and simulation techniques [1], [2].

The usage of traditional optimization methods is usually inefficient for solving multimodal or simulation-based problems. Therefore, methods designed for global optimization are important from a practical point of view. The problem of designing algorithms to compute global solutions is very difficult. In general there are no local criteria in deciding whether a local solution is a global one. During last decades, however, many theoretical and computational contributions helped to solve multiextreme problems arising from real-world applications [3]–[5].

In our paper we will present an integrated software environment, called global optimization object-oriented library (GOOL), which can be used to solve complex optimization problems. GOOL supplies the library of optimization algorithms for convex and nonconvex, unconstrained and constrained problems together with the graphical environment for supporting the considered problem definition and tools for dynamic, on-line monitoring of the computed results. The GOOL system integrates various functionalities, and can be successfully used in research, education and commercial applications. The preliminary version of the system was described in [6]. The currently available version is more advanced and has wider range of applications.

This paper is organized as follows. In Section 2 we will discuss the principle features of the global optimization algorithms. Next, we will describe organization, implementation and usage of our software platform GOOL, and numerical algorithms supplied in GOOL. Finally, the results of the application of solvers from the GOOL library to a price management problem will be presented and discussed.

2. Global Optimization Algorithms

Global optimization algorithms can be categorized into two groups: deterministic and stochastic, with respect to their implementation.

Deterministic algorithms are typically based on approximation techniques, approaches that adaptively perform partition, search and bounding, chaotic movement and tabu search. These methods usually require more or less access to global information about the problem. Many of them are guaranteed to find the global minimum (within some tolerance). A unified and insightful treatment of deterministic global optimization is provided in [3], [7], [8].

Stochastic algorithms are typically based on random search, adaptive search, biological inspired heuristics and metaheuristics. Heuristic stochastic methods are widely used in many industrial and scientific applications. These approaches are flexible, robust and less demanding of the problem properties. The main methodological and theoretical developments in stochastic global optimization, the basic principles and methods of global random search, Markovian and population-based random search and methods based on statistical models of multimodal functions are discussed in [9]. The evolutionary algorithms, genetic algorithms, genetic programming, learning classifier systems, evolution strategy, differential evolution, particle swarm optimization, and ant colony optimization, and other metaheuristics, such as simulated annealing, hill climbing, tabu search, and random optimization are elaborated in [4], [5], [10], [11].

Global optimization is generally complex and usually involves cumbersome calculations, especially when consider simulation-optimization case when we have to perform simulation experiment in every iteration of the algorithm. The restrictions are caused by demands on computer resources – central processing unit (CPU) and memory. The directions, which should bring benefits are:

- hybrid techniques that combines global and local algorithms, to solve the optimization problem;
- parallel computing where the whole task is partitioned between several cores, processors or computers.

Hybrid approaches can speed up the convergence to the solution. Parallel implementation allows to reduce the computation time, improve the accuracy of the solution, and to execute large program which cannot be put on a single processor [1], [12], [13].

3. GOOL: Software Environment for Global Optimization

In this section we present the design and implementation of GOOL and comparison of our project to the other existing tools for global optimization.

3.1. Related Works

Most of the existing libraries of optimization techniques focus on the problem of computing locally optimal solutions. However, recently a number of software packages with numerical solvers for global optimization have been developed, and can be find in the Internet. They support sequential and parallel programming. Publicly available implementations of interval analysis and branch-and-bound schemes are discussed in [14]–[16].

The goal of the COCONUT (continuous constraints – updating the technology) project [17] was to integrate the currently available techniques from mathematical programming, constraint programming, and interval analysis into a single discipline, to get algorithms for global constrained optimization. The authors of [18] report the results of testing a number of existing state of the art solvers using COCONUT routines on a set of over 1000 test problems collected from the literature. Solvers implementing various types of techniques for global optimization (deterministic and stochastic), i.e., interval methods, continuous branch and bound, multistart, genetic and evolutionary, tabu search and scatter search are provided in [15]. The Global World [19] is a forum for discussion and dissemination of all aspects of global optimization problems. It provides links to libraries of solvers and a library of academic and practical test problems.

3.2. GOOL Overview

The GOOL provides an integrated graphical software framework that can be used to solve the following very general problem:

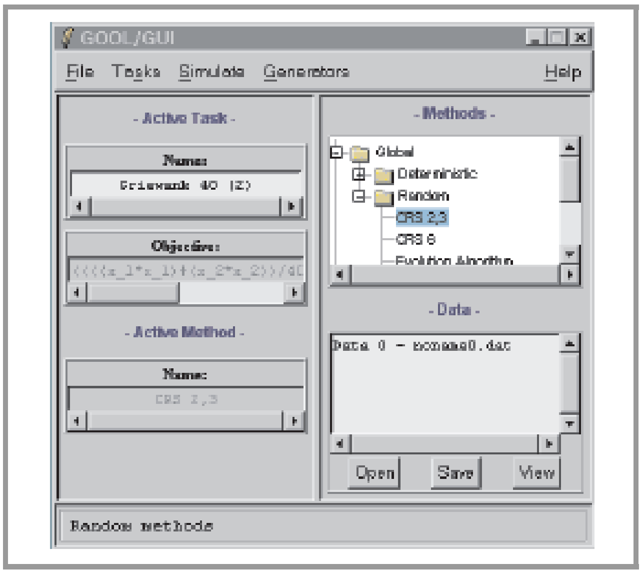
$$\min_{x\in\Re^n} f(x) \tag{1}$$

$$g_i(x) \leq 0, \qquad i = 1,\ldots,m,$$

where f and g_i are real-valued functions.

The GOOL supplies a library of deterministic and stochastic optimization solvers. When most of the available libraries for calculating the optimal solution provide tools only for commerce, research or educational purposes, the GOOL system integrates all these functionalities. The process of implementing a given application for GOOL is quite straightforward and convenient especially thanks to graphical user interface (GUI). The system provides tools for on-line monitoring of computation process and various presentation techniques.

Two different versions implementing two approaches to user-system interactions: GOOL/COM (batch) and GOOL/GUI (interactive) are supplied. GOOL/COM is dedicated to the complex optimization problems, where values of the objective function are calculated based on simulation (simulation-optimization scheme, [2]). In the case of simulation optimization the user's task is to provide the simulation model to evaluate the expected performance of the system to be optimized. It is assumed that solvers from the GOOL library provide decision variables and receive values of the objective function f and constraints g in Eq. (1) from the user application. Let the input files be called `task_file.tsk` and `methods_file.met`. Then, writing the command `gool_con task_file [methods_file]` at the command line, we call GOOL to solve the optimization problem defined in the file `task_file` using the optimization algorithm pointed in the file `methods_file`. The input file `task_file` contains the information related to the particular problem to be solved (problem dimension, objective function definition, its gradient and constraints) or the name of the user application (simulator). The selection of the solver is optional.



Name: Griewank 40 (2)

Objective: ((((x_1*x_1)+(x_2*x_2))/40

Name: CRS 2,3

Data 0 - sxname0.dat

Random methods

Fig. 1. GOOL/GUI: the main window.

The GOOL/GUI is the software framework for educational purposes and research (see Figs. 1 and 2). It supplies the graphical environment for optimization problem definition and calculation results presentation. The optimization prob-

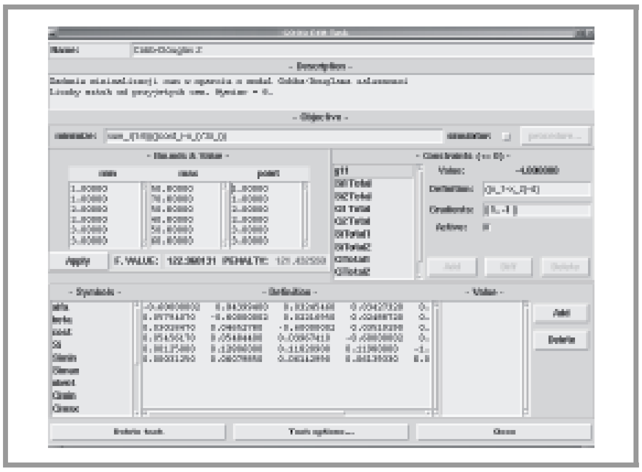


Fig. 2. GOOL editor: optimization problem implementation.

lem is defined using the GOOL editor. The GOOL symbolic expressions analyzer allows to enter quite complicated functions concerning such expressions like: pow, sin, sum, etc., and iterative expressions. The gradient is calculated if necessary. After starting the calculations the user can on-line employ the monitoring of the results.

3.3. System Architecture

The system consists of three components (Fig. 3):

- library of numerical methods,
- system kernel,
- graphical user interface.

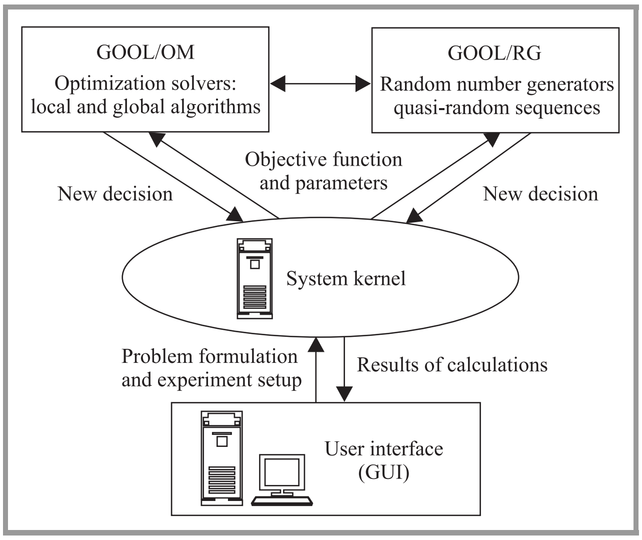


Fig. 3. The GOOL system architecture.

The core component is numerical library consisting of two sets of modules:

- GOOL/OM: optimization solvers,
- GOOL/RG: random number generators.

In addition the system facilities are provided in the form of four groups of services. These are:

1. User interface services, which provide a consistent user interface. The most important tasks of the user interface are as follows: supporting the process of defining a considered optimization problem, results visualization, providing communication with the user.
2. Calculation management services, which manage execution of a given solver.
3. Communication management services, which manage communications between calculation process and user interface.
4. Data repository services, which provide a store for all data objects: all defined options, parameters and calculation results.

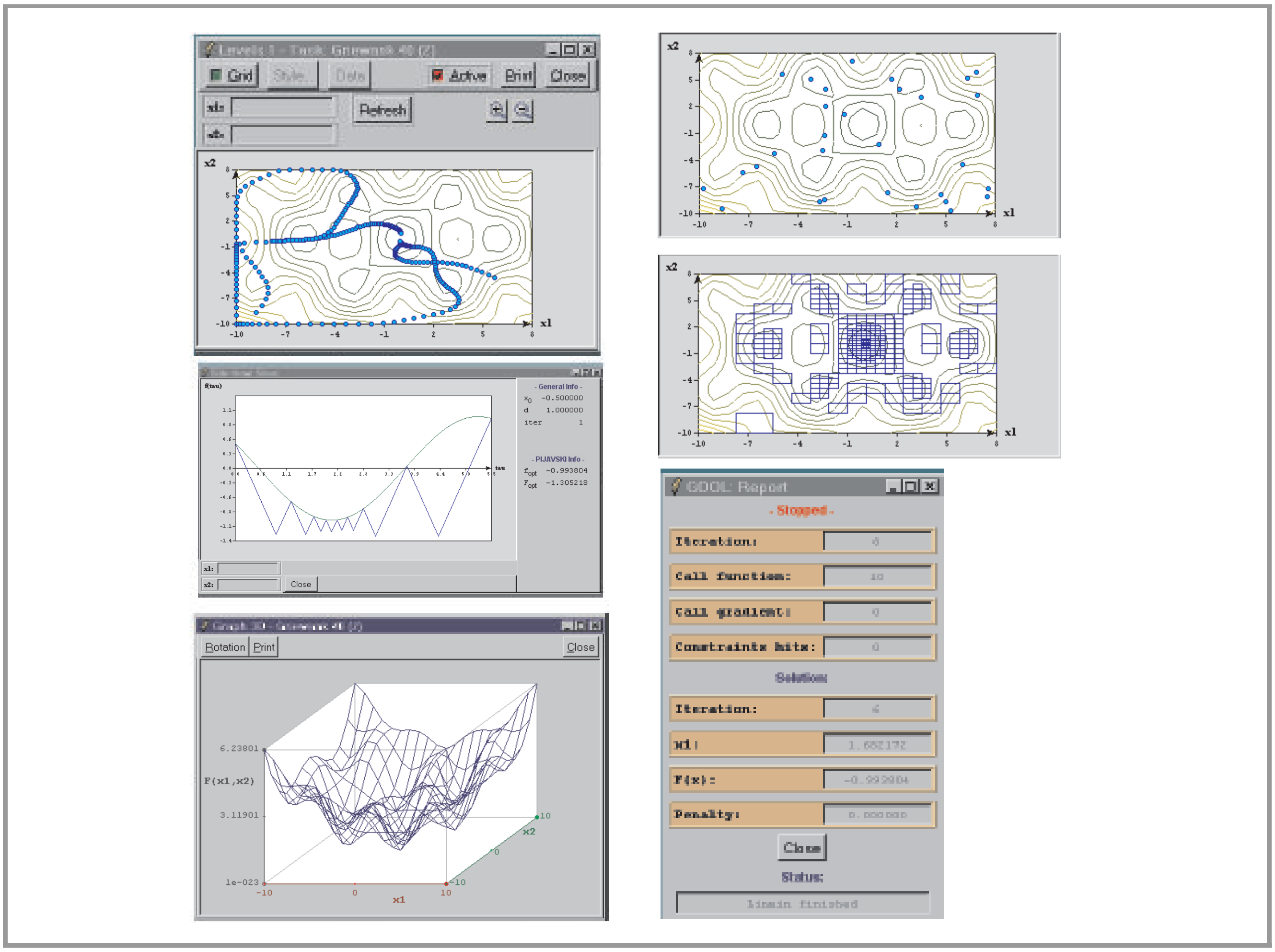
3.4. Algorithmic and System Options

Various algorithmic and system options are available to the user, all come with a default value so it is not necessary to modify any options. The ability to modify them, however, provides a great deal of flexibility. It is possible to change all parameters of the chosen solver, type of random generator or local search using graphical interface in GOOL/GUI or text file `methods_file` in GOOL/COM. Different termination criteria are provided: typical to each algorithm (if exists), convergence tolerance, maximum number of iterations or function evaluations. The results can be displayed every iteration or recorded into the disc file and displayed at any time.

3.5. Graphical User Interface

The GUI is organized in a set of windows. The setting windows are used to facilitate the configuration phase. The optimization problem is defined, an objective function and all constraints are entered.

The GOOL provides tools for dynamic, on-line monitoring of the computed results. The following graphical presentation techniques are available: 2D, 3D graphs, leaves of the function and a table of numbers (Fig. 4). The visualization of a multidimensional problem is achieved by displaying in the separate windows the leaves for each pair of variables, under the assumption that all other variables are fixed. The results presentation is organized in different ways, and is fitted to the optimization method (points, lines, grids). Multiple windows with the results for different range of data can be active during one run of the program. The changes of values of parameters typical to each algorithm can be graphically displayed too. The user can chose options of presentation (zoom, colors, results of many optimizations in one window, etc.). The detailed report of the results including the problem solution, number of iterations, number



- General Info -
x_0 -0.500000
d 1.000000
iter 1

- PIJAVSKI Info -
f_{opt} -0.993804
F_{opt} -1.305218

x1:
x2:

Iteration: 6
Call function: 10
Call gradient: 0
Constraints hits: 0

Iteration: 6
x1: 1.652172
F(x): -0.993804
Penalty: 0.000000

Status: linmin finished

Fig. 4. Results visualization.

of function and gradient evaluation and number of constraints violation is displayed after finishing the calculations.

3.6. GOOL Operation

The interaction with GOOL/GUI is organized as follows. At the beginning the user is asked to define the problem to be solved and select the optimization algorithm. Within the next step the user is asked to provide some information related to the considered method and calculation process if necessary. This information includes: parameters typical for the chosen algorithm, type of the stop criterion, maximal number of iterations, type of results visualization, etc. After completing the initial settings, GOOL starts the calculation engine. The user employs monitoring of the current situation. It helps him to assess the effectiveness of the chosen optimization algorithm. The calculations may be interrupted.

3.7. Implementation

The GOOL system is completely based on C++. All numerical methods – the optimization engine – and the higher-level activities, i.e., problem definition, parameters setting, results presentation, managing calculations and communication between the optimization engine and the user interface are implemented in uniform form as C++ classes. Two functionalities of GOOL, i.e., user interface and calculations are separated and can be easily modified. The hierarchy of classes implementing numerical solvers is natural and well defined (Fig. 5). Three fundamental clas-

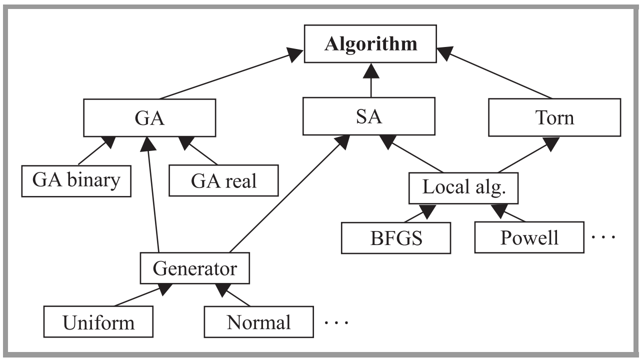


Fig. 5. GOOL: hierarchy of classes. Explanations: GA – genetic algorithm, SA – simulated annealing, BFGS – Broyden-Fletcher-Goldfarb-Shanno.

ses: `Task`, inserting the considered optimization problem to be solved, `Algorithm`, the basic class of all optimization methods and `Generator`, for random numbers generation are provided. The library can be extended by new methods developed by the user. Available software may be easily adopted, new techniques can be implemented applying classes defined in GOOL. The open design of the system architecture, and its extensibility to include other open source modules, was chosen in the hope that the system will be a useful platform for research and education in global optimization. The code is currently available for MS-Windows and Linux operating systems. The software is free available for researchers and students.

4. Library of Solvers

The numerical library consists of two parts: GOOL/OM and GOOL/RG. The GOOL/OM is a collection of different optimization solvers for calculating local and global minimum. GOOL/RG provides several random numbers generators.

4.1. Local Optimization

Several techniques for calculating local minimum of the performance index were implemented in GOOL. The following methods for one-dimensional search are available: golden section search, parabolic interpolation, one-dimensional search with first derivatives (Goldstein test). Two, well known nongradient methods in multidimensions [20] are available too: downhill simplex algorithm (Nelder-Mead) and direction set (Powell's) method.

4.2. Global Optimization

Deterministic and stochastic techniques are provided. Currently implemented are methods based on the approximation and branch-and-bound techniques, deterministic chaotic movement, clustering techniques, random search, heuristics and metaheuristics. The following variants are provided:

- Branch-and-bound (BB) for Lipschitz problems: uniform grid, few versions of non-uniform grids [3], [21]: Galperin's, Gourdin-Hansen-Jeaumard's, Meewella-Mayne's, and Pijavskij's algorithm of linear sub-approximations of the performance function, developed for one-dimensional problems.
- Chaotic movement: Griewank's algorithm (trajectory method) [22], [23].
- Clustering method developed by Torn [21], with different grouping techniques.
- Pure random search and three variants of population set based direct search methods controlled random search (CRS): CRS2, CRS3 and CRS6 as described in [13], [24].
- Simulated annealing (SA) as described in [25].
- Genetic algorithm (GA) using fixed-length binary strings for its individuals and evolutionary strategy (ES) with real-valued individuals [10], [11].

The available algorithms can be used to solve general constrained optimization problems. The constraints that cannot be handled explicitly are accounted for in the objective function using simple penalty terms for constraints violation. The reformulation of Eq. (1) is made inside the GOOL system:

$$\min_{x\in\Re^n} \left[f(x)+\Psi(x)\right], \qquad \Psi(x)=\mu\sum_{i=1}^{m}\max(0,g_i(x))^p . \quad (2)$$

The user can insert the value of parameters μ and p in Eq. (2).

4.3. Random Numbers Generation

Many heuristic algorithms provided in GOOL use random number generators to calculate a new decision. The large number of random generators have been developed over the last decades. Several procedures representing different types of generators are available in the library: uniform (two variants), normal (three variants), beta distribution, Cauchy distribution. Sequences of n-tuples that fill n-space more uniformly, than uncorrelated random points are called the quasi-random sequences [20]. That term is somewhat of a misnomer, since there is nothing "random" about quasi-random sequences – they are cleverly crafted to be, in fact sub-random. Three such sequences are available in GOOL: Halton, Sobol and Faure.

5. Case Study Results

5.1. Formulation of Price Management Problem

Several stochastic algorithms from GOOL library were compared. In this section we present the computational results obtained for prices optimization problem.

The considered case study was to calculate the optimal prices for products that are sold in the market. The goal was to maximize the total profit *PR*:

$$\max_x \left[PR=\sum_{i=1}^{n}\left(\frac{x_i}{(1+v_i)}-d_i\right)S_i(x)\right], \qquad (3)$$

where n denotes number of products exist (corresponding to n price decisions x_i), v_i and d_i are given constants corresponding to the market entities of VAT (value added tax) and cost per product, S_i are expected sales of product i within the considered period, assuming that prices of all products are fixed over this period. Several sales models can be found in the literature [26]. All these models describe market response on the price of jth product. We considered three of them.

Cobb-Douglas model. This model is following:

$$S_i(x) = \alpha_i \prod_{j=1}^{n} x_j^{\beta_{ij}}, \qquad (4)$$

where x_j denotes the price of product j, α_i is the scaling factor for sales of product i, β_{ij} is the elasticity of sales of product i with respect to the price of product j (β_{ii} is referred to as the direct elasticity and β_{ij}, $i \neq j$ is the cross elasticity).

Gutenberg model. The response function Eq. (4) is widely used but it does not capture some important effects, such as different market sensitivities to small and large price changes. Another sales model is formulated:

$$S_i(x) = a_i - bx_i + c_{1i}\sinh\left(c_{2i}(x_i - \overline{x}_i)\right), \qquad (5)$$

where a_i, b_i, c_{1i} and c_{2i} denote model parameters and $\overline{x}_i$ the average competitive price, i.e., price computed as the average of competitor prices taking into account their respective market share. The additional term can be added to this expression: $c_{3i}\sinh\left(c_{4i}(x_i - x_{i_0})\right)$, where x_{i_0} denotes the current price of the product i. The response function Eq. (5) belongs to the group of s-shaped models. The major difficulty is in fact that for some values of parameters such market response can involve multiextreme profit *PR* in Eq. (3), as presented in Fig. 6 (see [27] for details).

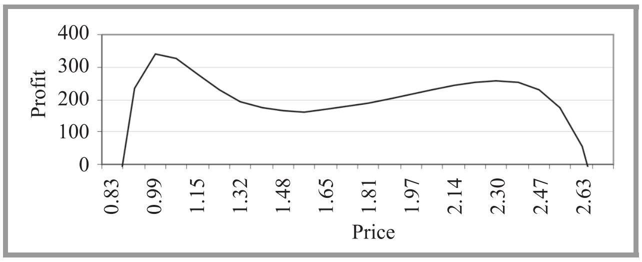


Fig. 6. Values of profit for different prices of a given product (model Eq. (5)).

Hybrid model. In the model Eq. (5) the cross-effects with other substitute or complementary own products are not included. The next considered model formulated in [28] combines functions Eqs. (4) and (5):

$$\begin{aligned} S_i(x) &= a_i - bx_i + \alpha_i \prod_{j=1}^{n} x_j^{\beta_{ij}} \\ &\quad + c_{1i}\sinh\left(c_{2i}(x_i - \overline{x}_i)\right) \\ &\quad + c_{3i}\sinh\left(c_{4i}(x_i - x_{i_0})\right), \end{aligned} \qquad (6)$$

where α_i, β_{ij}, $\overline{x}_i$ and x_{i_0} the same like in Eqs. (4) and (5), a_i, b, c_{1i}, c_{2i}, c_{3i}, c_{4i} model parameters. This model exhibits an s-shape and includes cross-effects.

Constraints. The following constraints for price, sale and cash of each product and for total sale and cash can be considered: $x_{i_{\min}} \leq x_i \leq x_{i_{\max}}$, $S_{i_{\min}} \leq S_i \leq S_{i_{\max}}$, $C_{i_{\min}} \leq x_i S_i \leq C_{i_{\max}}$, $TS_{\min} \leq \sum_{i=1}^{n} x_i \leq TS_{\max}$, $TC_{\min} \leq \sum_{i=1}^{n} x_i S_i \leq TC_{\max}$.

In listed constraints $x_{i_{\min}}$ and $x_{i_{\max}}$ denote minimal and maximal prices of product i, $S_{i_{\min}}$, $S_{i_{\max}}$ minimal and maximal sale, $C_{i_{\min}}$, $C_{i_{\max}}$ minimal and maximal cash, $TS_{\min}$, $TS_{\max}$ minimal and maximal total sale, and $TC_{\min}$, $TC_{\max}$ minimal and maximal total cash. In practice, usually prices of only some products are changed at anyone time. The following constraint restricts the number of prices, which can be modified

$$\sum_{i=1}^{n} \frac{\gamma(x_i - x_{i_0})^2}{1 + \gamma(x_i - x_{i_0})^2} \leq w, \qquad (7)$$

where γ and w are assumed parameters, x_{i_0} the current price of the product i.

5.2. Comparison of Market Response Models

The comparative study of all presented market response models was performed. The goal of the experiments was to calculate the optimal prices for fifteen products ($n = 15$ in Eq. (3)). The optimization model was defined using the GOOL graphical editor (see Fig. 2). All models parameters were randomly generated in ranges determined based on real historical data. The evolutionary strategy solver supplied in the GOOL library was used to solve the task.

The results – suggested prices of fifteen products – obtained for three presented market response models, taking into account only price bounds are depicted in Fig. 7. We can observe that the model Eq. (4) suggests the highest prices while the model Eq. (5) expresses less optimism suggesting lower prices. The results for Eq. (6) are between values obtained using Eqs. (4) and (5).

5.3. Comparison of Solvers

The goal of the second series of tests was to compare the efficiency of selected solvers from the GOOL library. The experiments were performed for historical data. Calculations were terminated after 100 iterations of each algorithm. The results obtained for 15 products, market response function Eq. (6) w.r.t. all listed constraints are compared in Table 1. The values collected in the adequate columns denote: *PR* – the total profit defined in Eq. (3), time – time of calculations in seconds.

Table 1
Simulation results for market response function Eq. (6)

Method	*PR*	Time [s]
CRS2	1252	91
SA	1286	96
ES	1281	11

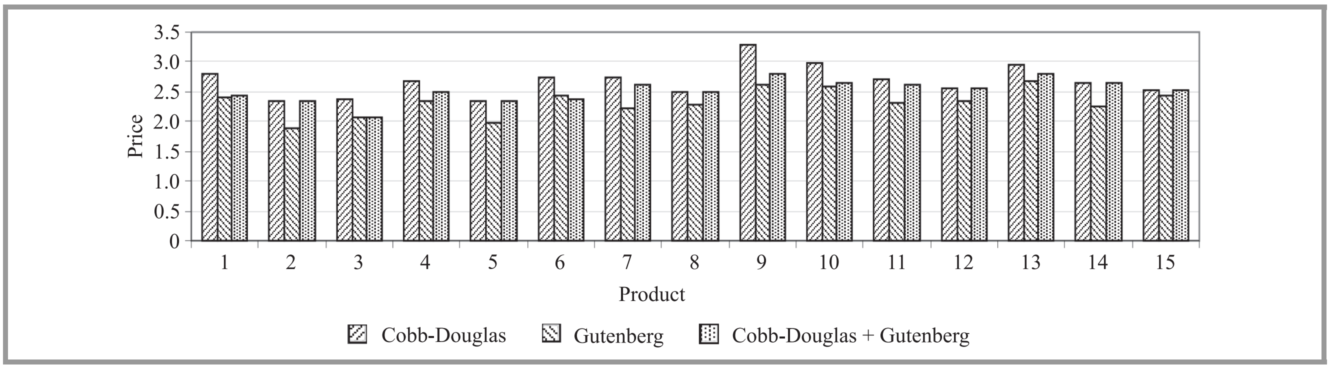


Fig. 7. Prices for different market response models.

The available numerical results indicate that ES and SA methods give better solution than CRS method. The best result was obtained using SA but the time required to compute solution was longer than ES method. Attempts to solve the considered problem using other solvers provided in GOOL (clustering method, branch-and-bound and chaotic movement) failed. The feasible solution was not found or the computation time was unacceptable long. The conclusion to be drawn is that heuristics such as ES, although quite simple are efficient and robust for many real-life optimization problems.

Finally, two versions of controlled random search methods described in [13] and [24], i.e., CRS2 and CRS6 were compared. CRS are population set based random search algorithms. The basic random search consists of three main steps: generate the initial set of points, transform the population, and check the assumed stopping condition. Several versions of CRS methods related to different strategies of new trial points calculations were developed. CRS2 is the simplest one. CRS6 is much more advanced – it uses quadratic interpolation and random numbers generation from the β distribution to calculate new trial points. The weakness of all CRS methods is the way in which the constraints of a type $g_i(x) \leq 0$ are handled. The infeasible points are simply rejected from further consideration. The suggested approach is to use penalty terms for constraints violation Eq. (2).

The optimization results of price management problem Eq. (3) with sales model Eq. (4) and all listed constraints, considering CRS2 and CRS6 methods are presented in Tables 2 and 3. The experiments were performed for several sets of historical data, containing various groups of products offered in supermarkets. Prices of 15, 31 and 53 products were calculated. Each solver was executed five times, the assumed accuracy was 10^{-4}. The results obtained for modified objective function Eq. (2) were compared with those obtained for the standard approach that discards infeasible points (Table 2). The values collected in tables denote: n – number of products, PR – average total profit, time – average time of calculations in seconds, f_{call} – average number of the objective function Eq. (3) evaluations.

The results presented in Tables 2 and 3 indicate that the CRS2 algorithm is very fast but only gives an approximate solution, even in the case when the penalty function is used (see Table 2). The CRS6 method provides better results with respect to CRS2 but the time required to compute a solution was longer than the CRS2 method.

Table 2
Total profit *PR* in case of two approaches
to infeasible points (15 products)

Method	Discarding infeasible points	Penalty for constraints violation
CRS2	1215.95	1237.45
CRS6	1241.27	1241.27

Table 3
Comparison of the fastest and the most
accurate methods

n	Best PR	Algorithm	f_{call}	PR	Time [s]
15	1241.27	CRS2	24002	1235.41	1.98
		CRS6	39392	1241.27	3.16
31	830.75	CRS2	69562	805.76	23.39
		CRS6	122298	830.71	33.64
53	544.65	CRS2	76169	526.03	23.85
		CRS6	285849	544.65	75.99

As a conclusion the following strategy is proposed: in cases when accuracy of the solution is the crucial the CRS6 method with the discarding of infeasible points are suggested; when it is crucial that the problem is solved quickly the CRS2 method with the penalty function should be used.

6. Summary and Conclusions

In this paper a brief description of the software platform GOOL for complex systems optimization was made. GOOL was design to be powerful, effective, flexible, and easy to use software for optimization. It is suitable to solve different optimization problems and can be successfully used for global minimum calculating. The user-friendly interface allows to perform the numerical experiments in the effective manner both for research and education. The open design of the system architecture, and its extensibility to include new solvers make GOOL be a useful platform for global optimization. The current version of GOOL can support researchers and engineers during the design and control of real-life complex systems in the sense of decision automation. In our future research we plan to extend our system to multiobjective optimization to provide the tool that will support interactive optimization process.

Acknowledgments

This work was partially supported by Ministry of Science and Higher Education grant NN514 416934.

References

[1] E. Niewiadomska-Szynkiewicz, "Symulacja komputerowa w analizie i projektowaniu złożonych systemów sterowania", Warsaw, Warsaw University of Technology Press, 2005 (in Polish).

[2] J. C. Spall, *Introduction to Stochastic Search and Optimization*. New Jersey: Wiley, 2003.

[3] R. Horst and P. M. Pardalos, *Handbook of Global Optimization*. Dordrecht: Kluwer, 1995.

[4] Z. Michalewicz and D. B. Fogel, *How to Solve it: Modern Heuristcs*. New York: Springer, 2000.

[5] T. Weise, *Global Optimization Algorithms: Theory and Application*, e-book, 2009 [Online]. Available: http://www.it-weise.de/projects/book.pdf

[6] M. Publicewicz and E. Niewiadomska-Szynkiewicz, "GOOL – global optimization object-oriented library", in *Proc. KAEiOG'2003, Conf.*, Łagów, Poland, 2003, pp. 173–181.

[7] A. C. Floudas, *Deterministic Global Optimization: Theory, Methods and Applications*. Dordrecht: Kluwer, 1999.

[8] A. Neumaier, *Complete Search in Continuous Global Optimization and Constraint Satisfaction*, Acta Numerica. Cambridge, Cambridge University Press, 2004, pp. 271–369.

[9] A. A. Zhigljavsky and A. Zilinskas, *Stochastic Global Optimization*. Springer Optimization and Its Applications. New York: Springer, 2007.

[10] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*. New York: Springer, 1997.

[11] R. Schaefer, *Foundations of Global Genetic Optimization*. Berlin-Heidelberg: Springer, 2007.

[12] A. Karbowski and E. Niewiadomska-Szynkiewicz, "Obliczenia równoległe i rozproszone", Warsaw, Warsaw University of Technology Press, 2001 (in Polish).

[13] W. L. Price, "Global optimization by controlled random search", *JOTA*, vol. 40, no. 3, pp. 333–348, 1983.

[14] K. Holmqvist and A. Migdalas, "C++ class library for interval arithmetic in global optimization", in *State of the Art in Global Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Dordrecht: Kluwer, 1996.

[15] "Solver technology – global optimization" [Online]. Available: http://www.solver.com/technology5.htm

[16] S. Tschoke and T. Polzer, "Portable parallel branch-and-bound library: PPBB-Lib, user manual, library version 2.0", University of Paderborn, Germany, 1999.

[17] "COCONUT – continuous constraints updating the technology" [Online]. Available: http://www.mat.univie.ac.at/∼neum/glopt/coconut

[18] A. Neumaier, O. Shcherbina, W. Huyer, and T. Vinko, "A comparison of complete global optimization solvers", *Math. Programm.*, vol. 103, no. 2, pp. 335–356, 2005.

[19] "Global World Forum" [Online]. Available: http://www.gamsworld.org/global/index.ht

[20] W. H. Press, S. A. Tukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, The Art of Scientific Computing. Cambridge, Cambridge University Press, 1992.

[21] A. Torn and A. Zilinskas, *Global Optimization*, LNCS, vol. 350. Berlin: Springer, 1989.

[22] A. O. Griewank, "Generalized descent for global optimization", *J. Opt. Theory Appl.*, vol. 34, no. 2, pp. 11–39, 1981.

[23] J. W. Rogers and R. A. Donnelly, "A search technique for global optimization in chaotic environment", *JOTA*, vol. 61, no. 1, pp. 111–121, 1989.

[24] M. M. Ali and C. Storey, "Modified controlled random search algorithms", *Int. J. Comput. Math.*, vol. 53, no. 3–4, pp. 229–235, 1994.

[25] A. Dekkers and E. Aarts, "Global optimization and simulated annealing", *Math. Programm.*, vol. 50, no. 1–3, pp. 367–393, 1991.

[26] H. Simon, *Price Management*. North-Holland: Elsevier, 1989.

[27] M. Dygas and E. Niewiadomska-Szynkiewicz, "Optymalna wycena produktów i usług – modele, oprogramowanie i eksperymenty symulacyjne", Int. Rep. ICCE WUT, no. 03-17, Warsaw, 2003 (in Polish).

[28] K. Malinowski, "PriceStrat 4.0 Initial Research Paper", KSS Int. Doc., Manchester, 2000.

Ewa Niewiadomska-Szynkiewicz received her Ph.D. in 1996, D.Sc. in 2006 from the Warsaw University of Technology. She works as a Professor of control and computation engineering at the Warsaw University of Technology. She is the Head of the Complex Systems Group. She is also an associate professor at the Research and Academic Computer Network (NASK), and the Director for Research of NASK since 2009. She is the author or co-author of three books and over 90 journal and conference papers. Her research interests focus on complex systems modeling and control, computer simulation, global optimization, parallel calculations and computer networks. She was involved in a number of research projects including three EU projects, coordinated the Groups activities, managed organisation of a number of national-level and international conferences.

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

e-mail: ewan@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Michał Marks – for biography, see this issue, p. 41.

Paper

Query Optimization in Teradata Warehouse

Agnieszka Gosk

Abstract—The time necessary for data processing is becoming shorter and shorter nowadays. This thesis presents a definition of the active data warehousing (ADW) paradigm. One of the data warehouses which is consistent with this paradigm is teradata warehouse. Therefore, the basic elements of the teradata architecture are described, such as processors parsing engine (PE) and access module processor (AMP). Emphasis was put on the analysis of query optimization methods. There is presented the impact of a primary index on the time of query execution. Furthermore, this paper shows different methods of optimization of data selection, data joins and data aggregation. All these methods can help to minimize the time for data processing. This paper presents experiments which show the usage of different methods of query optimization. At the end some conclusions about different index usage are included.

Keywords—active data warehouse, query optimization, teradata.

1. Introduction

The time of data processing is important nowadays. There are popular solutions which improve this time, for example, OLAP systems, streaming databases. There is also a new solution – active data warehousing (ADW), which is not used as often as the systems mentioned before.

The ADW paradigm is related to data warehouse, which is updated as fast as possible. ADW allows minimization of the time between events and decisions which are made in connection with this event. Therefore, such decisions are much more valuable. The primary objectives of ADW are to decrease the time of decision-making, as well as to enhance the reliability of these decisions [1].

The reliability of decisions can be increased thanks to basing them on current data. Therefore, in data warehouse which is consistent with the ADW paradigm, data should be updated as fast as possible [2]. At the moment of appearance of any event (modification of the data in a source system) the data warehouse should be updated. It is possible by introducing a mechanism of triggers, which after the appearance of an event that is meeting certain conditions, update the data warehouse [3].

A rapid response to a query is possible through optimization of queries. The overview of selected query optimization methods is the objective of this document.

2. Theoretical background

2.1. Teradata Architecture

The teradata architecture is very specific. It is presented in Fig. 1 [4].

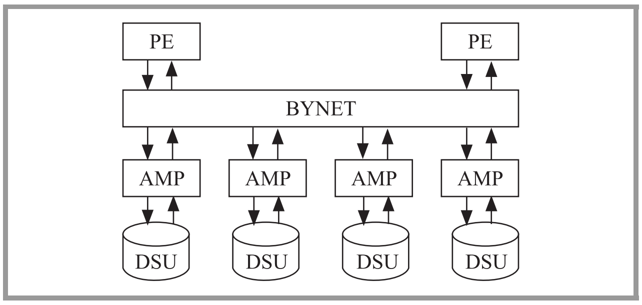


Fig. 1. The architecture of teradata warehouse.

The BYNET is a high-speed network element. It is used to transfer data between parsing engine (PE) and access module processor (AMP).

The PE is a virtual processor. It is responsible for communication between client application and the database (receives a request from the client applications and returns response rows to the requesting client). When PE receives a query, it checks the session parameters (manages session) and divides the query into steps. Then it controls the step execution that is performed by the AMPs. The PE has a few elements, which are described below.

The parser checks if a query, which was sent by a client application, is written correctly. It checks its syntax and whether the user has appropriate rights to all objects, which were used in that query. The optimizer chooses the best method of query execution. For example it can choose a sequence of table joins. The best method of the query execution is presented as a tree and is sent to the generator. The generator converts the tree, which was sent by the optimizer, into steps and sends all the steps to the dispatcher. The dispatcher sends steps of the query to the appropriate AMPs. Then it controls the execution of all the steps and the sequence in which they are executed. Some steps can be executed parallelly, but there are steps that can be executed only after finishing other steps.

The AMP is a virtual processor. It controls a specific disk subsystem – virtual disk. The AMP manages its own disk subsystem and sets the response rows on the basis of its own disk subsystem. It can execute aggregation, sorting,

joins. All data transformations are executed according to the steps, which were sent by the PE.
The disk storage unit (DSU) is a physical disk subsystem – virtual disk. It is managed by one and only one AMP [4].

In teradata warehouse there are several types of indexes:

– primary index,

– secondary index,

– join index,

– hash index.

It is important to build indexes, because they can dramatically improve the time of data processing. The description of each type of indexes is given below.

2.1.1. Primary Index

It is the most important index, because it has to be in each table. When a primary index is created, the database does not build any additional table, which can store values of that index. If a primary index is not created on a table, then the database creates it. On the basis of values of the column set that define the index, the hash value of each row is calculated. This value determines on which AMP the mentioned row is going to be kept. Therefore, when a search for data with a specific index value is performed, the database hashes this value. On the basis of this hash value the database knows on which AMP appropriate data is stored. Only one AMP software is searching for the requested data. So the entire table does not have to be scanned.
There are two types of primary indexes:

– unique primary index (UPI),

– nonunique primary index (NUPI).

In a table with a unique primary index each value of primary index has to be unique [4].

2.1.2. Secondary Index

Secondary indexes are not required. They do not affect the data distribution. Like a primary index, a secondary index can be:

– unique,

– nonunique.

A secondary index can improve the time of query execution, when a table with a defined secondary index value is searched. Additionally, a unique secondary index forces uniqueness of the index values.
When a secondary index is built on a base table, a subtable is created. This subtable stores secondary index values, secondary index hash values and hash values of the primary index of each row. Therefore, when a search for data with a defined secondary index value is preformed, the database hashes this value. The AMP, which was indicated by the secondary index hash value, searches for the appropriate row in the subtable. In this row there is the hash value of the primary index. This last hash value indicates the AMP which stores the requested row from the base table. Finally, the indicated AMP searches for the requested row [4].

2.1.3. Join Index

A join index can be defined in one or more tables. When the join index is built on a base table/tables, a subtable is created. In this subtable there is a copy of some data from the base table/tables or a subset of base table columns. A query can be executed accessing the index (subtable) instead of joining and accessing the base tables. Generally, join indexes can improve the time of data processing [4].

2.1.4. Hash Index

A hash index can be compared with a join index and a secondary index. Like the join index defined on one table, the hash index can redistribute rows from the base table across the AMPs. Like a secondary index each row of the hash index has a pointer to an appropriate row from the base table [4].

Summarizing, it is known that the primary indexes influence data distribution. Therefore, they can improve the time of all operations. The time of data selections can be improved by primary index, secondary index or hash index. The data joins can be executed faster thanks to primary indexes, secondary indexes, hash indexes or join indexes. The join indexes also improve the time of data aggregations.

But it is not known how strong various types of indexes can improve the time of different operations. It is difficult to say how much disk storage various indexes can occupy or how many costs they cause. In the next part of this paper some experiments are presented, which give some answers.

3. Experiments

Experiments were performed on the same server, with the following parameters:

- Dual Core AMD Opteron
  Processor 880
  2,39 GHz
- 2,00 GB RAM

The Microsoft Windows Server 2003 Standard Edition system was installed on the server.
Experiments were executed on the Teradata Warehouse 8.1Demo. In this version of the teradata system there are only 2 AMP processors and 1 PE processor available. In the DEMO version BYNET element does not exist and disk space is limited to 4 GB.

Two tables were prepared and used to carry out tests. A definition of these tables is presented below.

```
CREATE TABLE CLIENTS (
    client_id INTEGER NOT NULL,
    name VARCHAR(20),
    surname VARCHAR(20),
    street VARCHAR(20),
    home_no VARCHAR(8),
    city VARCHAR(20),
    tariff_plan VARCHAR(15),
    status CHAR(1),
    phone_type VARCHAR(18),
    phone_number INTEGER,
    activation_date DATE);
```

```
CREATE TABLE CLIENT_CONNECTIONS(
    client_id INTEGER NOT NULL,
    connection_type VARCHAR(1),
    connection_direction VARCHAR(2),
    count INTEGER,
    volume INTEGER);
```

The CLIENTS table stores basic data about clients, who are active. The CLIENT_CONNECTIONS table stores data about connections, which were performed by active clients. In the experiments described, various primary indexes and added indexes are defined in the tables. These indexes are presented in experiment descriptions.

For each table four sets of data were prepared. In each set there was a different number of clients and a different number of rows. Tables 1 and 2 present information about the CLIENTS and CLIENT_CONNECTIONS tables.

Table 1
The CLIENTS table statistics

Data set	Number of clients	Number of rows	Size[B]
A	500 000	500 000	43 822 080
B	1 000 000	1 000 000	86 585 344
C	1 500 000	1 500 000	129 408 512
D	2 000 000	2 000 000	172 274 688

Table 2
The CLIENT_CONNECTIONS table statistics

Data set	Number of clients	Number of rows	Size[B]
A	500 000	2 141 184	77 952 000
B	1 000 000	4 282 709	155 574 784
C	1 500 000	6 425 485	232 721 920
D	2 000 000	8 567 631	310 469 632

The problem of query optimization is not new. There are many papers concerning this. Many of the optimization methods show how to write a query. Disappointingly, in teradata warehouse these methods are not effective. But there are methods of query optimization, which rely on index usage. These methods can be used in teradata warehouse. In this paper these methods are presented.

3.1. Influence of Data Distribution on Query Execution Time

As it was said, a primary index is the most important index which influences data distribution. In this section experiments which show how data distribution determines the time of the execution of various types of queries, are presented.

Experiment 1

Firstly, there was checked the primary index affected on the time of selection of all rows from the CLIENTS table. In this experiment a following query was executed:

```
SELECT *
FROM CLIENTS;
```

The skew factor shows how equally data is distributed across the AMP processors. The bigger this factor value is, the more unequally data is distributed. During this experiment the primary index of the CLIENTS table was changed, so the skew factor of the CLIENTS table changed. In Fig. 2, there are line graphs which show the query execution time in seconds as a function of data set size for different values of the skew factor.

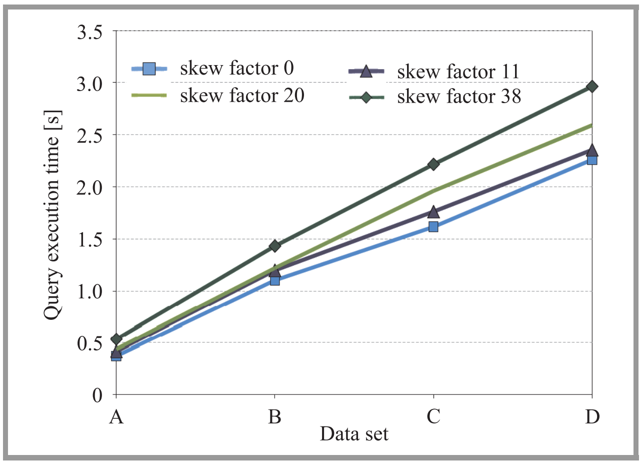


Fig. 2. Influence of the primary index on the time of data selection.

In the figure it is shown that the query execution time increases with growing number of rows. For the same set of data, the query execution time is slower for the CLIENTS table with bigger skew factor values (about 25% for data set D). Therefore, it must be remembered that unequal data distribution influences the time of data selection significantly.

Experiment 2

This experiment checks how data distribution influences the time of data join. For this test a following query, which chooses number of different connections from CLIENT_CONNECTIONS table for active clients from CLIENTS table, was selected:

```
SELECT C.client_id, CC.connection_type,
     CC.connection_direction, CC.count
FROM CLIENTS C,
     CLIENT_CONNECTIONS CC
     C.client_id = CC.client_id;
```

This query was performed on various sets of data. In the CLIENTS and CLIENT_CONNECTIONS tables there were different primary indexes (they were created on different columns sets). Once the skew factor for the CLIENTS table was 0 and once 38. Results are shown in Fig. 3, there are line graphs which show the query execution time in seconds as a function of data set size for different indexes available in the database. The line graphs are marked according to description:

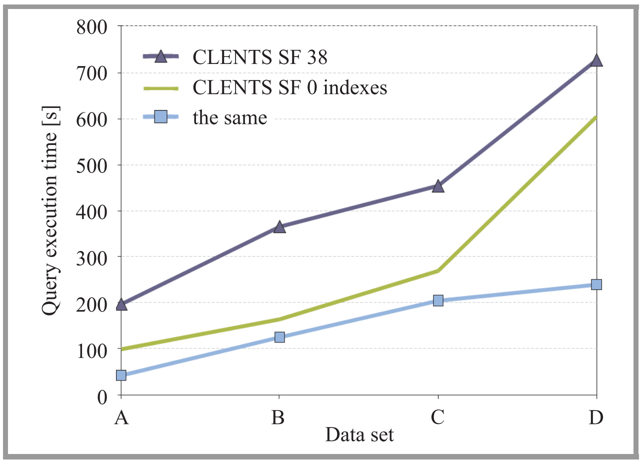


Fig. 3. Influence of the primary index on the time of data joins.

- The same indexes. In the CLIENTS and CLIENT_CONNECTIONS tables there were the same primary indexes. They were defined on the client_id column.
- CLIENTS SF 0. In the CLIENTS and CLIENT_CONNECTIONS tables there were various primary indexes. The skew factor of the CLIENTS and CLIENT_CONNECTIONS tables was 0.
- CLIENTS SF 38. In the CLIENTS and CLIENT_CONNECTIONS tables there were different primary indexes. The skew factor of the CLIENTS table was 38, and the skew factor of the CLIENT_CONNECTIONS table was 0.

It can be seen, that the time of data join depends on a value of the skew factor. When only one table has a high value of the skew factor the time of query execution is getting worse. In the case when skew factor of CLIENTS table is 38, the time of data processing is even 2 times greater than in the case when skew factor of CLIENTS table is 0. The time of data joins is the best when the tables, which are joined, have the same primary indexes.

3.2. Different Methods of data Selection Optimization

Experiment 3

This experiment examines how indexes can influence the selection of one row from a table.
For the needs of this experiment a following query, which chooses one row from the CLIENTS table, was prepared:

```
SELECT *
FROM CLIENTS
WHERE phone_number = 300001019;
```

During this experiment indexes, which were created in the phone_number column of the CLIENTS table, were changed. Figure 4 presents the experiment results – the query execution time in seconds as a function of data set size in a form of different line graphs. These experiments were performed when different indexes were available. The description of graphs is presented below:

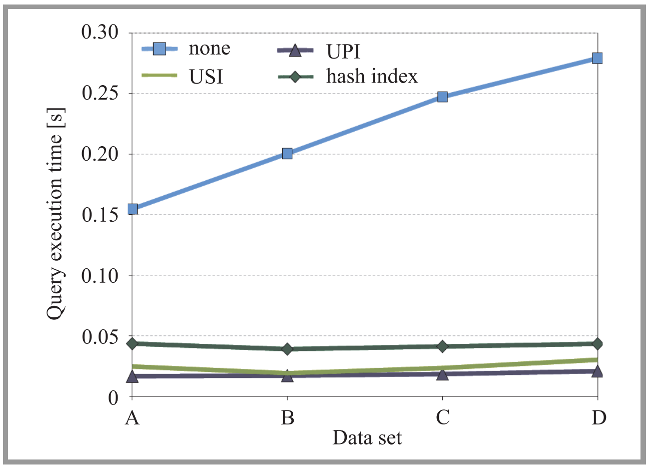


Fig. 4. Influence of various indexes on the time of data selection.

- UPI – a unique primary index of the CLIENTS table is defined on the phone_number column,
- NONE – a primary index of the CLIENTS table is defined on the client_id column, on the phone_number column there is no index,
- USI – a unique secondary index of the CLIENTS table is defined on the phone_number column, the CLIENTS.client_id column – is the primary index,

- Hash – a hash index of the CLIENTS table is defined on the phone_number column.

The worst results are received when there is no index on the phone_number column. The best results are received when the primary index is created on the phone_number column. When database uses USI or hash index the results are similar to results when the database uses UPI. In Fig. 4 it can be seen that whatever any index is used: hash, secondary or primary, the time of selection of one row from the table is not dependant on the number of rows in this table.

It must be remembered that USI and hash indexes cause additional costs, they increase the time of table updating and they use additional storage space. In Table 3 there is presented the time of 2 000 000 rows insertion and deletion from the CLIENTS table when additional indexes are available in the database.

Table 3
The time of 2 000 000 rows insertion and deletion from the CLIENTS table (data set D), when additional indexes are available in the database

Operation	NONE	USI	Hash index
Insert [s]	1	9	603
Delete [s]	1	1	238

Secondary and hash indexes on the CLIENTS.phone_number column require additional storage space (for data set D) - 68 MB and 84 MB. Therefore, it is better to create the secondary index, because it influences the time of data insertion and deletion from the CLIENTS table less than the hash index and it occupies less disk space than the hash index.

Experiment 4

The next experiment checks how different indexes influence the selection of many rows from one table.

To carry out this test, a following query, which chooses clients from the CLIENTS table who use PT_1 tariff plan, was prepared:

```
SELECT *
FROM CLIENTS
WHERE tariff_plan = 'PT_1';
```

During this experiment the number of available tariff plans was changed. A different number of clients used plan 'PT_1', therefore, the query which is presented above, returns a different number of rows. The results of this experiment are presented in Fig. 5, there are line graphs which show the query execution time in seconds as a function of number of rows, which are returned, for different indexes available in the database. The line graphs are described as it is presented below:

- NONE – a primary index of the CLIENTS table is defined on the client_id column, on the tariff_ plan column there is no index.

- NUPI – a nonunique primary index of the CLIENTS table is defined on the tariff_plan column.

- NUSI – a unique primary index of the CLIENTS table is defined on the client_id column and a nonunique secondary index of this table is defined on the tariff_plan column

It can be seen that the best results are received when in the CLIENTS.tariff plan column the nonunique primary index is created. The worst results are received when in the same column there is no index.

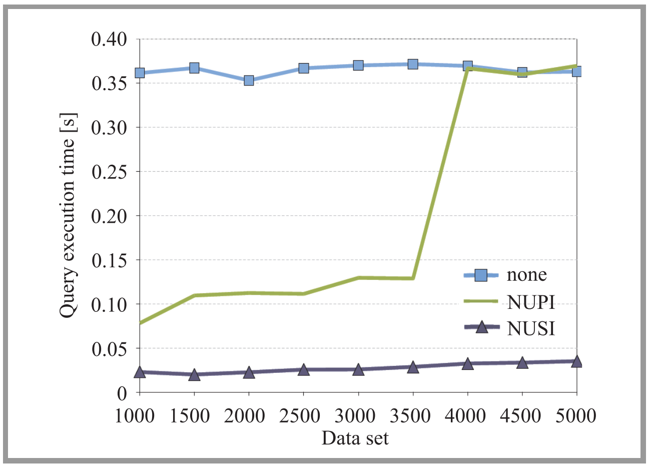


Fig. 5. Influence of different indexes on the time of selection of many rows from the table.

An interesting line graph is received when on the CLIENTS.tariff_plan column a nonunique secondary index is created. When a query returns 3500 records or less the time of query processing is quite good, but when a query returns more than 3500 records, the results are the same as in the case when on the CLIENTS.tariff_plan column there is no index. It is so because when a query returns more than 3500 records it is more efficient to retrieve rows from the base table than from the secondary index. When command EXPLAIN is used to check how a query is executed, two different explanation are returned. One, in the case when a query returns 3500 records or less, the other when a query returns more than 3500 records.

NUSI occupies about 20 MB of disk space and it influences the time of data insertion and deletion from the CLIENTS table slightly. Therefore, when a primary index cannot be created on a column set, which is used for data selection, it can be replaced by a secondary index.

3.3. Different Methods of Optimization of Data Joins

This section presents how different type of indexes can influence the time of data join. It is checked by the experiment which is presented bellow.

Experiment 5

To execute this experiment the following query was prepared, which calculates the number of types of connection directions for voice connections for active clients who have PT_1 tariff plan:

```
SELECT C.client_id,
   COUNT(DISTINCT CC.connection_direction)
FROM CLIENTS C,
   CLIENT_CONNECTIONS CC
GROUP BY C.client_id
WHERE C.client_id = CC.client_id
   AND CC.connection_type = 'V'
   AND C.tariff_plan = 'PT_1';
```

The query which was presented above, was executed using different types of indexes. The results are presented in Fig. 6, where the line graphs show the index influence on the time of data join according to description:

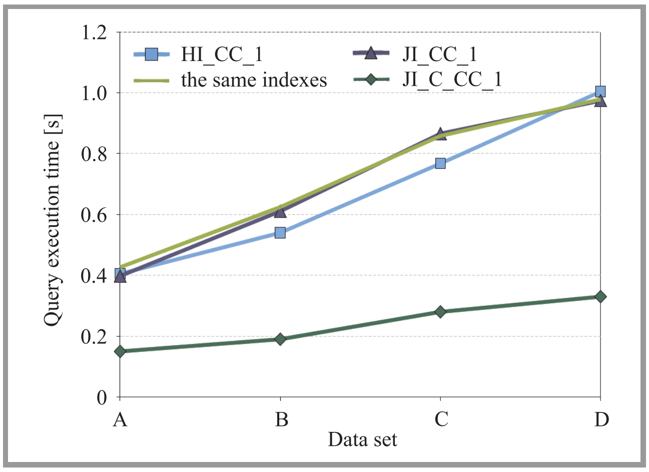


Fig. 6. Influence of different indexes on the time of data join.

- JI_CC_1 – a join index JI_CC_1 is defined on the CLINET_CONNECTIONS table;
- JI_C_CC_1 – a join index JI_C_CC_1 is defined on the CLIENTS and CLIENT_CONNECTIONS tables;
- HI_CC_1 – a hash index HI_CC_1 is defined on the client_id column of the CLIENTS_CONNECTIONS table;
- THE SAME PI – In the CLIENTS and CLIENT_CONNECTIONS tables there are the same primary indexes (on the client_id column).

Definition of join indexes JI_CC_1 and JI_C_CC_1 is presented below.

```
CREATE JOIN INDEX JI_CC_1 AS
SELECT *
FROM CLIENT_CONNECTIONS
PRIMARY INDEX(client_id);
```

```
CREATE JOIN INDEX JI_C_CC_1 AS
SELECT C.client_id, CC.connection_direction
FROM CLIENTS C,
   CLIENT_CONNECTIONS CC
WHERE C.client_id = CC.client_id
   AND CC.connection_type = 'V'
   AND C.tariff_plan = 'PT_1'
PRIMARY INDEX(client_id);
```

When the JI_CC_1 join index, the hash index or the same primary indexes on the CLIENTS and CLIENT_CONNECTIONS tables are used, the time of query processing is similar. In each case data in the JI_CC_1, HI_CC_1 and the CLIENT_CONNECTIONS table is distributed by the client_id column. When the query is executed, data is read directly from the JI_CC_1, HI_CC_1 or the CLIENT_CONNECTIONS table in the different cases and it is joined with the CLIENTS table in each case. In joined tables data is distributed by the same key, therefore, data is joined in AMPs.

When on the CLIENTS and CLIENT_CONNECTIONS tables there are different indexes, the time of query processing is 20 minutes for data set D. It is so because data has to be redistributed across AMPs during join.

The best query execution time is received when join index is defined on the CLIENTS and CLIENT_CONNECTIONS tables (JI_C_CC_1). In this case the two tables do not have to be joined, data can be read directly from JI_C_CC_1. But indexes like JI_C_CC_1 do not have as wide usage as indexes like JI_CC_1.

It must be remembered that the JI_C_CC_1 index influences the time of data insertion and deletion from the CLIENTS and CLIENTS_CONNECTIONS tables. Indexes JI_CC_1 and HI_CC_1 influence the time of data insertion and deletion only from the CLIENTS table. When in the database the JI_C_CC_1 index is defined, the time of 8 567 631 rows insertion into the CLIENT_CONNECTIONS table is about 3 minutes, when in the database index JI_CC_1 or HI_CC_1 is defined this time is about 20 minutes. It is so because the time of data insertion or deletion from the table is dependent on the number of rows from this table, which are inserted into the index, which is defined on this table. Therefore, the more rows in the index, the longer time of data insertion and deletion from the table.

The size of indexes JI_C_CC_1, JI_CC_1 and HI_CC_1 is 308MB, 690KB and 119 MB. It could be critical when database storage is limited.

3.4. Different Methods of Data Aggregation Optimization

This section presents a method of data aggregation optimization. The experiment performed is presented below.

Experiment 6

To show how different indexes influence the time of data aggregation, a query was prepared, which summarizes the number of connections and volume of these connections for clients from the CLIENT_CONNECTIONS table:

```
SELECT client_id,
    SUM(count) as count,
    SUM(volume) as volume
FROM CLIENT_CONNECTIONS CC
GROUP BY client_id, ;
```

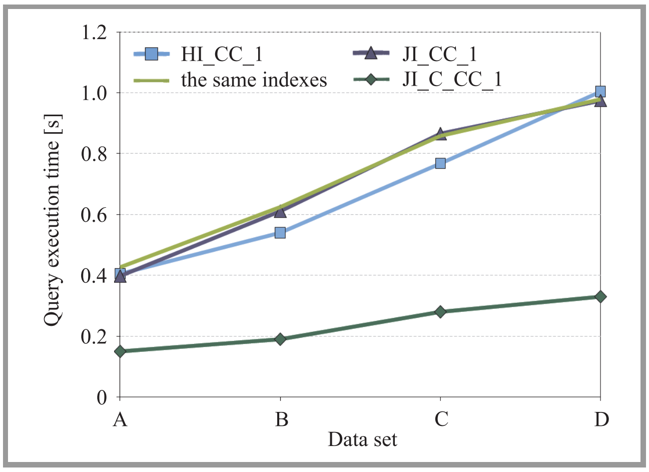


Fig. 7. Influence of different indexes on the time of data aggregation

The query was executed when various indexes were available in the database. The results are presented in Fig. 7 and particular line graphs show the time of query processing in seconds as a function of data set size, when different indexes are available:

- NONE – in the CLIENT_CONNECTIONS table there is a primary index defined on the column set: client_id, connection_type, connection_direction;
- AI_CC_1 – in the database there is available the AI_CC_1 index;
- AI_CC_2 – in the database there is available the AI_CC_2 index;
- PI – in the CLIENT_CONNECTIONS table there is a primary index defined on the client_id column.

A definition of AI_CC_1 and AI_CC_2 indexes is presented below:

```
CREATE JOIN INDEX AI_CC_1 AS
SELECT client_id, connection_type,
    SUM(count), SUM(volume)
FROM CLIENT_CONNECTIONS
GROUP BY client_id, connection_type;
```

```
CREATE JOIN INDEX AI_CC_2 AS
SELECT client_id, SUM(count), SUM(volume)
FROM CLIENT_CONNECTIONS
GROUP BY client_id;
```

When in the database there are not available any join indexes, the time of query processing is not good. It can be improved when a join index is created or when the primary index is changed. When the primary index is created on the same column as the column which groups data in the query, which is presented above, the time of query execution is quite short. The reason for this behavior is because data aggregation is performed in AMPs.

When in the database the AI_CC_2 index is available, the time of query processing is promising. It is because data is read directly from the index, and it does not need to be aggregated. While in the database there is available the AI_CC_1 index, the time of query processing is worse. Data is read directly from the index, but it has to be additionally aggregated. However the AI_CC_2 index has wider usage than the AI_CC_1 index. The AI_CC_1 can be additionally used to query which groups data by client_id and connection_type. The AI_CC_1 index occupies more disk storage than the AI_CC_2 index.

4. Conclusions

As it was said the primary index is the most important index. It has influence on the time processing of all operations. The time of data join is shorter when joined tables have the same primary indexes. The time of data aggregation is shorter when a table has the same primary index as the grouping column in a query. However, as the experiments have shown, when the primary index is chosen, the most important is the skew factor value of the table. The smaller it is, the better the processing time of all operations is. Furthermore, the primary index should be adapted to executed joins and aggregation.

When there is need to optimize time of data selection on the basis of column set, we can choose the primary index on this column set, a secondary index or a hash index. However, the secondary index and the hash index increase the time of insertion and deletion of data from a table. Most often the hash index has a bigger maintenance cost than the secondary index.

During the data join, there can be used: a join index, a hash index or the primary index can be changed. Different indexes cause different additional costs. Most frequently, the hash index causes lower costs than the join index. Data aggregation can be optimized by the join index or the primary index. The join index which has aggregated data can help to avoid aggregation during query execution.
The above-mentioned conclusions were drawn on the basis of experiments, which were presented in the previous sections. However, in different conditions costs caused by indexes can change. Therefore, it should be remembered that when indexes are chosen, it is most important to calculate additional costs caused by this indexes. It must be known what is more important the processing time improvement or the database space size. Then the best indexes can be chosen.

References

[1] S. Brobst and J. Rarey, “The five stages of an Active Data Warehouse evolution”, *Teradata Magazine Online*, 2001 [Online]. Available: http://www.ncr.com/online_periodicals/brobst.pdf

[2] M. Gonzales, “Getting Active”, *DB2 Mag.*, iss. 1, Q1, 2005 [Online]. Available: http://www.dbmag.intelligententerprise.com/story/showArticle.jhtml?articleID=59300861

[3] E. Kanana and M. Farhi, “Enhancing data preparation processes using triggers for active datawarehousing”, in *Proc. Int. Conf. Data Mining*, Las Vegas, USA, 2006, pp. 153–160.

[4] Teradata Documentation, Database Design, pp. 335–561, Introduction to Teradata. NCR Corporation, 2005.

Agnieszka Gosk received the M.Sc. degree in computer science from the Warsaw University of Technology (WUT), Poland, in 2009. She had been employed by the National Institute of Telecommunication in Warsaw till 2009. She is currently working in the area of data warehousing for a telecommunication operator. Her scientific interests include: data mining, modeling and decision support.
e-mail: agnieszkagosk@gmail.com

Paper

Solving Support Vector Machine with Many Examples

Paweł Białoń

Abstract—Various methods of dealing with linear support vector machine (SVM) problems with a large number of examples are presented and compared. The author believes that some interesting conclusions from this critical analysis applies to many new optimization problems and indicates in which direction the science of optimization will branch in the future. This direction is driven by the automatic collection of large data to be analyzed, and is most visible in telecommunications. A stream SVM approach is proposed, in which the data substantially exceeds the available fast random access memory (RAM) due to a large number of examples. Formally, the use of RAM is constant in the number of examples (though usually it depends on the dimensionality of the examples space). It builds an inexact polynomial model of the problem. Another author's approach is exact. It also uses a constant amount of RAM but also auxiliary disk files, that can be long but are smartly accessed. This approach bases on the cutting plane method, similarly as Joachims' method (which, however, relies on early finishing the optimization).

Keywords—concept drift, convex optimization, data mining, network failure detection, stream processing, support vector machines.

1. Introduction

The application of optimization methods in data analysis, especially in telecommunications, yields optimization problems with a very specific structure. To the author's opinion, this specificity will have to make deep changes in the optimization science itself, by forcing the algorithm designers to work with unusual circumstances and requirements.

We shall exemplify this claim with the case of linear classification of points in $\mathbb{R}^N$, each preassigned to one of two classes: A or B. We shall deal with the optimization problem encoding the linear classification task, called support vector machine (SVM) problem. This problem will be precisely formulated later. Now it suffices to say that the problem of linear classification consists in finding a hyperplane in $\mathbb{R}^N$ that properly (or as properly as possible) separates these points into the classes.

Looking at the SVM problems that are nowadays analyzed, we notice that many of them are obtained automatically. This is very common to the telecommunication applications. For a very simplified example, each "point" can represent a state of a telecommunication network measured with the simple network management protocol (SNMP), with coordinate values representing, e.g., the traffic in particular arcs of the network, particular elements of the connection matrix, error parameters, etc. The two classification classes could be the proper state of the network or a failure, and the classification hyperplane, for some training points, pre-assigned to these classes by a teacher, could be further used in automatic failure detection.

This example shows two specific structural properties of the data:

1. There may be very many classification points. For example, this will happen if the SNMP data come at regular time intervals like tens of seconds and are collected through a long period, perhaps several months. In the resulting optimization, it will be possible that the random access memory (RAM) exhausts with all this data, so we may be not able to store the optimization problem in RAM.
2. The data may be very dense, resulting with optimization problems that are unusually dense for the optimization standards. Usually we hope for some level of sparsity of optimization problems claiming that the the input data must be in some way verified by a human and that he cannot conceive too many nonzero numbers. Now, however, the situation becomes different: the data is not produced by a human, like a modeler cooperating with the optimization expert but produced automatically. And it is not surprising that each sample is relatively dense in our example: the traffic volume in a particular arc of the network is usually nonzero at any moment.

   Having a large, dense optimization problem is a very untypical case for a common imagination of a specialist at optimization.

The author believes the above two features can be also present in many other applications in which the data is obtained automatically at regular time bases, e.g., as the log of the behavior of customers of telephony subscribers, bank clients, supermarket clients, medical sensor data, etc.

Stream processing. The extreme case of dealing with long streams of input points is the case of *stream processing*. Stream processing (see [1]) is a general data-mining concept, relevant to problems with data that can be aligned in a stream of similar items, like records in a database. In linear classification, we can have a stream of preassigned points. The algorithm for solving a problem with such data has the stream processing character if it uses memory constant in the stream length (number of items in the stream). This means that the each incoming portion of the data from the input stream has to be processed in a sense on-line, i.e., the algorithm can, for example, update some partial stream statistics with this portion of data but cannot remember all the data read so far. We can think of stream processing like a new name for an old concept of “ideal processing algorithm”.

Stream processing is *very* unusual in optimization. Almost all optimization algorithms assume they have random access to particular parameters defining the optimization problem, i.e., it difficult to predict which parameter the algorithm will read in its next iteration. Also, the algorithm can return to some parameter so far read, i.e., it can read one parameter several times. Thus all the parameters defining the problem must be constantly accessible. If they do not fit all together in RAM but, for example, fit in a disk file, we still can think of building a smart *oracle* that communicates the required parameter to the algorithm, cleverly (fast) navigating through the file: we shall show such a solution. In the case of stream processing we cannot even have a long file, and this situation requires a completely new approach to solving the optimization problem.

Concept drift. Some methods of processing long streams are able to take into account the phenomenon of *concept drift*. To explain this phenomenon it is convenient to think of the input stream as of infinite. The algorithm solves its problem periodically, and each time the problem instance is defined with the portion of the stream from the beginning to the last portion of data read. It may happen that the incoming data slowly change in time because of the reality or the phenomena described by the data also slowly change. For example, the number of user of the computer network we probe increases, and this changes the traffic characteristics. This is called concept drift and the solution of our problem, like the separating hyperplane, must also slightly evolve in time. Thus to take into account the concept drift, our algorithm must be first of all capable of giving periodic solutions with the portions of the input data stream “from the beginning up to now”. Moreover, we often impose some gradual forgetting of older data: the older the data is, the less it weights in the definition of the current problem instance. Of course, a precise definition of weighting would have to be written, dependent on the particular problem being solved. Still it is reasonable to assume that our memory is far too low to store all the “new” part of the input data stream in.

In our critical analysis in which we use results obtained by Joachims in [2], the author’s own result from [3] and a new author’s concept.

2. Linear Support Vector Machine Problem

Linear support vector machine problem [4], [5] is a certain formalization of the problem of finding a hyperplane separating as well as possible points (training examples) in $\mathbb{R}^N$ that have been preassigned to two classes A or B each. There are many variants of the way of detailed posing this problem. We shall consider the variant with an affine hyperplane and inexact separation.

We have n training examples a_j, $a_j \in \mathbb{R}^N$ for $j = 1,..,n$ each either of class A or class B. We look for a separating rule of the form

$$\omega^\top x \geq \gamma, \tag{1}$$

where $x \in \mathbb{R}^N$ is a variable while $\omega \in \mathbb{R}^N$, $\gamma \in \mathbb{R}$ are the classifier parameters.

To obtain ω and γ we solve the following linear support vector machine optimization problem:

$$\begin{aligned} &\underset{\omega\in\mathbb{R}^N,\, \gamma\in\mathbb{R},\, y\in\mathbb{R}^n_+}{\text{minimize}} \quad \frac{1}{2}\|\omega\|^2 + Ce^\top y \qquad (2)\\ &\text{subject to}\\ &-d_j \cdot (a_j{}^\top \omega - \gamma) - y + 1 \leq 0\,, \quad \text{for } j = 1,\dots n\,. \end{aligned}$$

Each d_j is either -1 – if example a_j is of class A or 1 – if example a_j is of class B.

The optimal ω and γ of this problem yield the separating Ineq. (1) that can be used to classify any point $x \in \mathbb{R}^N$ during the classifier working phase: if the rule is satisfied for x, then x is classified to class A, else it is classified to class B.

The obtained separation hyperplane tries to conceive two phenomena depicted in Fig. 1: separation violation and separation with a margin.

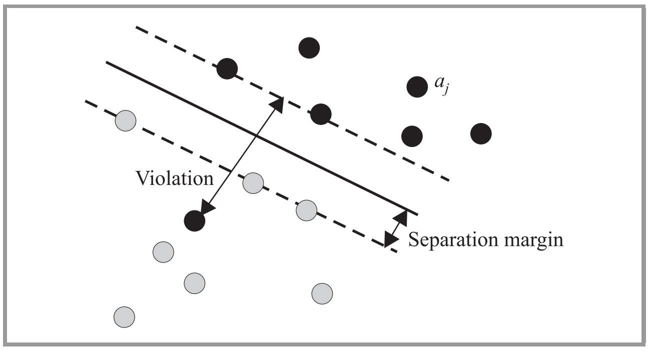


Fig. 1. Separation margin and separation errors. The training points are black and grey, indicating their belonging to one of the two classes.

A *separation margin* is obviously needed to avoid errors in classification. The points given to the classifier are distributed similarly but not identically as the training points. In turn, we allow that little training points be misclassified by the separation hyperplane, first because the problem may be not exactly linear separable, some training points may be distorted or in other way invalid, or there is too little of them to reasonably require the exact separation of them, scarifying other properties of the separation hyperplane.

The variables y_j represent the separation violations of particular training points. It can be shown that the separation equals to $1/\|\omega\|$ – since we do not want to go into details of the scaling present in problem (2), we can refer the reader to [4] for the proof.

Instead of maximizing $1/\|\omega\|$, we can minimize $\|\omega\|^2$, which is easier. Having said this, we can see what the goal function of problem (2) expresses: we tend to minimize the total separation violation and maximize the separation margin, the weight controlling the compromise is C.

This paper will deal mainly with two cases of size of this problem:

1. Number n of examples is large. This is a simpler case.

2. The RAM used by the algorithm is at most constant in n. This is a more difficult case of stream processing.

As opposed to n, we shall assume the number N of features is at most moderate. Otherwise, having in mind also the density of the problem, the problem would become too difficult even for most approaches discussed in this paper.

3. Approaches with Cutting Planes Oracles, Generation of Constraints and Cutting Planes

For problems with many constraints it is natural cutting plane methods connected with the oracle module that knows the problem instance and generates proper cuts.

Two of the approaches discussed here – [2] and [3] are concretizations of this idea. They differ slightly in the details of the formulation of the SVM problem but here they both can be described in terms of problem (2). Both the approaches assume the training examples are stored simultaneously in memory, so the oracle can return to some example.

Reformulation of the problem. First, we write an equivalent form of problem (2), in order to get rid of numerous decision variables y_j:

$$\underset{\omega\in\mathbb{R}^N,\,\gamma\in\mathbb{R},\,z\in\mathbb{R}}{\text{minimize}}\ \frac{1}{2}\|\omega\|^2 + Cz \tag{3}$$

subject to

$$\sum_{j\in\{1,\ldots,n\}} \max(0, -d_j\cdot({a_j}^\top\omega-\gamma)+1) - z \leq 0\,.$$

A further, redundant reformulation is:

$$\underset{\omega\in\mathbb{R}^N,\,\gamma\in\mathbb{R},\,z\in\mathbb{R}}{\text{minimize}}\ \frac{1}{2}\|\omega\|^2 + Cz \tag{4}$$

subject to

$$\sum_{j\in I} \max(0, -d_j\cdot({a_j}^\top\omega-\gamma)+1) - z \leq 0\ ,\ \text{for } I\in 2^{\{1,\ldots n\}}\,.$$

The equivalence of the formulations comes from the nonnegativity of the terms summed. Because of this nonnegativity, all the constraints in problem (4) are implied by the constraint in problem (3).

There is a huge number, $2^{\{1,\ldots,n\}}$ of constraints in problem (4). Certainly, all of them have their representation in memory. Instead, some constrained unsatisfied by an algorithm iterate x^k is generated by the oracle that is given x^k (if all the constraints are satisfied at x^k, the oracle returns a proper cut based on the gradient of the goal function at x^k). The gradient of this constraint defines a cut in our algorithm.

The reason of introducing redundant constraint is to accelerate the algorithm. Computing the gradient of a constrained in which the summations runs only over some subset I of $\{1,\ldots,n\}$, can be computationally easier than computing the gradient of the constraint in problem (3), which requires summing over $\{1,\ldots,n\}$.

Finding an unsatisfied constrained does not mean to try all the $2^{\{1,\ldots,n\}}$ constraints. A constraint with a bigger set I can be certainly obtained by an update of a constraint of a lower I. Thus the oracle needs only a single loop. In its consecutive iterations,the current I is enhanced by a new j. If we go up to the situation $I = \{1,\ldots,n\}$ having not found any unsatisfied constraint, we know there is no unsatisfied constraints (since we add positive numbers). Then the oracle can return a cut based on the gradient of the goal function.

Solving the reformulated problem. We shall compare the 2 approaches.

In [3] the problem (4) is tackled as follows.

1. The problem is solved with the Nesterov analytic cutting plane method with a penalty term [6].

2. The input data, defining the problem instance, is stored in a disk file, as it is too big to fit in RAM.

3. The oracle reads the file but since reading files is slow, the way the oracle navigates through the file is smart. Namely, it involves two accelerating mechanisms
   (a) The first one is the already defined incremental construction of constraints within the oracle
   (b) In the late stages of an optimization run, the above mechanism is inefficient, since near the solution, most of the problem constraints are satisfied, so one call of the oracle usually involves reading nearly or exactly n training points. But near the solution, the iterate does not move too much between iteration. So, instead of explicit checking violation of constraints by x^k we can assess this violation using the knowledge whether the respective constraint was violated by some earlier iterate, say x^{k-s}. The details of the assessment are described in [3]. It leads to the necessity of bucketing the input file, a certain surrogate of sorting this file due to some quantity.

It is interesting that in navigating our file we had to use a language characteristic for more traditional data processing or for databases rather than to optimization, e.g., we used bucketing. In the author's opinion this may be very indicative for the future of optimization science, that will be faced long streams of automatically generated data.

4. The unconstrained subproblems from the Nesterov method are solved directly (in primal).

In the approach of Joachims [2]:

1. Problem (4) is solved with the Kelley cutting plane method [7].

2. The subproblems from the Kelley method are transformed to their duals before being solved.

3. The k subproblems dual (the dual of the subproblem in kth iteration of the Kelley method) is a dense problem with about k variables. Also, k is the number of cuts made so far.

4. There are several interesting features of the Joachms' approach:

   (a) The most astonishing is its linear complexity in both N and n under given accuracy demand. This will be discussed later.

   (b) The method does not invert large matrices. The only matrix that might have to be invert may be the hessian matrix in some particular method solving a subproblem from the Kelley method; this hessian, however, is dense and of the size about $k \times k$ while even the largest k is assumed to be at most moderate in the algorithm, as discussed later.

Effectiveness under large number of examples. The story about how Joachims achieves his annoying linear complexity in both n and N is very meaningful and illuminating.

The author of this paper has made some experiments with the Joachims' solver. What quickly stroke was a very low (loose) default accuracy setting for this solver.
It turned out that the exceptionally good complexity in N and n is obtained at a cost of the rather quick dependence of the number of cuts (cutting plane iterations) on accuracy and the weight C. The number of the cutting plane method iterations is assessed as

$$\frac{8CR^2}{\varepsilon^2}, \tag{5}$$

where R is the radius of the set of a_js, ε is the solution accuracy in terms of the goal function. A comment is owed to the influence of C. The higher C, the less is the resulting separation margin and separation violations are penalized more. This makes the problem obviously harder, thus a larger number of necessary iterations of the method is not surprising.

An important conclusion is that a great gain in the speed of processing of long files of automatically generated data is to loosen the demands on accuracy.
A question that arise is whether the accuracy wanted by Joachims is sufficient in practice. The answer is not clear. A reasoning conducted in [3] says it is not enough. Simply, Joachims assumes that the number of iterations done by the cutting plane method will be low, even lower than the number of features N. Then the dense subproblem with an approximately $x \times k$ hessian is solvable within a reasonable time[1]. However, the approach becomes problematic when we see that the number of iterations of the cutting plane method is equal to the number of cuts generated during the optimization run. Our geometric intuition says that to properly isolate the solution in $\mathbb{R}^N$ by cuts we need rather of the rank of N cuts. In [3] we consider the following example.
Assume the number of cuts generated by the algorithm of Joachims is at least DN where D is a positive constant. Then only the last iteration of this algorithm costs

$$O\left(\theta(DN) + nD^2N^3\right). \tag{6}$$

This result is obtained under a reasonable assumption that solving a minimal optimization-state-of-art cost of solving kth subproblem and the costs of transformation to the dual. The above cost is already not linear in N.

Table 1
Comparison of the solvers' reaction to increasing C, problem `covtype`, $n = 523293$, $N = 54$, default accuracies

C	0.1	10	1000
Time – author [s]	1572	1510	1453
Time – Joachims [s]	384	4708	2739

Table 2
Comparison of the solvers' reaction to increasing C, problem `biology`, $n = 131320$, $N = 74$, $C = 0.1$; ε, is the accuracy setting for the Joachims' solver

e	0.0001	0.01	1000
Time – Joachims [s]	(> 2 hours)	118	287

However, the experiments with data coming from the practice do not support this theoretical reasoning. Neither do the experiments in [2] nor the experiment the author of this paper did in [3]. In the later experiments, a similar pattern of the solvers of Joachims and the solver of the author of this paper occurred. With the default settings and relatively low C. Increasing C and/or decreasing the solution, we quickly stuck the Joachims' solver, while the authors' solver, though maybe slower, obtained the solution (see the sample Table 1 and Table 2 for the experiments on bench-

[1] Moreover, it is still solvable in the case we have excluded from the scope of this paper, when N is big, which is the case definitely too difficult to the [3] approach, as the nondifferentiable Nesterov method will not work well with many variables of the subproblem.

mark examples from KDD04[2]). The authors' solver use the same very tight, default accuracy, measured in terms of distance from the solution rather than in terms of goal function.

Most interesting was that this increasing C or decreasing accuracy did not lead to the increase in the quality of the obtained classifier, measured by the accuracy of the classifier, i.e., the percent of well classified testing examples.

Thus the approach of Joachims – obtaining high efficiency, algorithm simplicity (e.g., no need to invert large matrices) consciously sacrificing some accuracy is the potential way of solving optimization problems with large, dense, automatically generated data. Optimization specialists should take this way into account and most attention should go to research on what accuracies are acceptable in practice.

Also, we see by the solution from [3] that optimization will have to borrow some language and solutions from databases or from more traditional data processing domains (e.g., sorting).

4. Approach for the Stream Case

Both the above solutions were semi-tools for the question of large data streams. They both allow returning to a particular training example, thus are not feasible for streams. We present below an idea for proceeding in the such a case.

For streams in optimization, the most natural approach is to make a model of the optimization problem that fits in memory constant with the stream length. We shall not go beyond this obvious approach, unlike, say, the ambitious approach in [8], in which we see some stream attitude in this that an optimization algorithm is itself essentially stream: new portion of data cause an update in the solution. However, the accuracy of the solution obtained in [8] is not great and the algorithm actually has an option to return to items previously read from the stream.

We shall use problem (3). Note that the most difficult in this problem is the sum in its constraint, which makes the constraint not storable in memory constant in n.

Note, however, that we have the nondifferentiable function $\max(0,\cdot)$ in the components of this sum. However, if we replace $\max(0,\cdot)$ with a polynomial $\phi : \mathbb{R} \mapsto \mathbb{R}$ the constraint will be storable in such RAM.

So, we can solve problem (3) in the following steps.

1. Reformulate SVM problem (2) as

$$\underset{\omega\in\mathbb{R}^N,\gamma\in\mathbb{R}}{\text{minimize}}\ \frac{1}{2}\|\omega\|^2 + C\sum_{j=1,\dots,n}\phi\left(-d_j\cdot({a_j}^\top\omega-\gamma)+1\right).$$

2. We approximated $\max(0,\cdot)$ in the constraint of problem (3) by a polynomial $\phi(\cdot) : \mathbb{R} \mapsto \mathbb{R}$, say, for example, of order 3.

   The constraint of the approximate problem is easily storable in RAM constant in n, since each function under sum is of the form $\psi : \mathbb{R}^{N+1} \mapsto \mathbb{R} \equiv \phi(w^\top x)$, where $w \in \mathbb{R}^N$ and is representable as:

$$\begin{aligned}\psi(x) \ = \ & \sum_{k=1}^{N+1}\sum_{l=1}^{N+1}\sum_{m=1}^{N+1} T^3_{k,l,m}x_kx_lx_m + \sum_{k=1}^{N+1}\sum_{l=1}^{N+1} T^2_{k,l}x_k \\ & + \sum_{k=1}^{N+1} T^1_k x_k + T^0.\end{aligned}$$

   To sum such vectors we need to respectively add the tensors defining particular components. So effectively we need one 3-dimensional tensor, one matrix, one vector and the constant. All of these objects have sizes dependent only on N.

3. Solve the approximate problem

Certainly, the open problem is how to choose the approximating polynomial so that the perturbance of the original problem is acceptable in practice. Also, perhaps more attention will be directed to operating on dense matrices/tensors, i.e., approximations with forcing some element values to zeros.

5. Conclusions

The conclusions from this work are following.

1. The practice yields new challenges to the science of optimization that have a potential of substantially change the research in optimization

   (a) The data created automatically can form very long streams, that are not storable in RAM and even force the algorithm to have a stream character, i.e., the memory usage constant in the stream length.

   (b) Such automatically generated data can be dense, resulting in dense optimization problems. We are used to the situation in which a human validates all the nonzero coefficient defining an optimization problem, thus there may be not too many of them. With an automatic generation, this argument is not valid.

2. The work of the Joachims shows that the solution to both the large size of the stream and the density of the data is to cleverly use some relaxations in the required accuracy. Experiments shows, some surprisingly, that so obtained solutions can be useful in practice. Even better effects (stream optimization) can be obtained by reformulating the whole optimization problem, not only the solution tolerance. Thus, further research should be directed to formally describing how a practical problem suffers from its formalization as an optimization problem being approximated.

[2] http://www.kdd.ics.uci.edu

References

[1] E. Ikonomovska, D. Gorgevik, and S. Loskovska, "A survey of stream data mining", in *Proc. 8th Nat. Conf. Int. Particip. ETAI 2007*, Ohrid, Republic of Macedonia, 2007, pp. I6-2.

[2] T. Joachims, "Training linear SVMs in linear time", in *Proc. ACM Conf. KDD 2006*, Philadelphia, USA, 2006, pp. 217–226.

[3] P. Białoń, "A linear Support Vector Machine solver for a huge number of training examples", *Control Cybern.* (to appear).

[4] D. R. Musicant, "Data mining via mathematical programing and machine learning". Ph.D. thesis, University of Wisconsin, Madison, 2000.

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

[6] Yu. Nesterov, "Complexity estimates of some cutting plane methods based on the analytic barrier", *Math. Program.*, vol. 69, 149–176, 1995.

[7] J. Kelley, "The cutting plane method for solving convex programs", *J. Soc. Ind. Appl. Mathem.*, vol. 8, pp. 703–712, 1960.

[8] A. Bordes and L. Bottou, "The Huller: a simple and efficient online SVM", in *Machine Learning: ECML-2005, Lect. Notes Artif. Int.* Springer, pp. 505–512.

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

Planning System for Multi-Agent Based Reconfigurable Fixtures

Wojciech Szynkiewicz

Abstract—This paper describes a concept of the planning system for self adaptable, reconfigurable fixtures composed of mobile locators (robotic agents) that can freely move on a bench and reposition below the supported part, without removing the part from the fixture. The main role of the planner is to generate the admissible plan of relocation of the mobile agents. A constrained nonlinear optimization problem is formulated to find the optimal locations for supporting heads.

Keywords—fixture planner, multi-agent system, optimization.

1. Introduction

The fixture planning system is an important element in computer aided process planning systems [1]. A fixture is a device for locating, constraining, and adequately supporting a workpiece during a manufacturing operation. Fixturing, like grasping seeks arrangements of contacts that restrict the possible motions of a given part. An important factor in fixture design is to optimize the fixture layout, i.e., positions of mobile locators, so that workpiece deformation due to clamping and machining forces is minimized [2], [3]. In this paper we consider the manufacturing process consists of milling (contouring) of thin-sheet aluminium parts for aircrafts and automotive bodies. Workpiece deformation is unavoidable due to its elastic nature, and the external forces impacted by the clamping actuation and machining operations. When severe part displacement is expected under the action of imposed machining forces, supports are needed and they should be placed below the workpiece to prevent or constrain deformation.

The existing fixtures for thin-walled workpieces like sheet-metal parts with complex surface geometries are:

- large mould-like fixtures,
- modular flexible fixture systems (MFFSs),
- single structure flexible fixture systems (SSFFS).

The fixtures traditionally used in manufacturing of thin-sheet metal parts are large moulds reproducing the shape of the skin to be supported, but this type of fixture is part specific and not reconfigurable. Usually, the mould surface is equipped with vacuum suction chambers and channels for holding the skin.

The MFFSs can be further classified on the basis of their adjusting mechanism:

- Partially reconfigurable with limited number of supports that can be manually relocated.
- Adjusted by separate devices, e.g., robot manipulators.
- Self-reconfigurable with a matrix of support elements with embedded actuators (in each locator/clamp).

It should be noted that all such fixtures still require some human intervention to reconfigure. Various MFFSs have been proposed [4]–[6], but their usage for thin-walled parts fixturing is rather limited. Since fixturing requirements vary during the different machining operations required on a single part, it becomes necessary to reposition the supports, interrupting the production process. MFFSs can be adapted to various parts but their initial cost is often high while configuration is complex and time consuming.

One way to avoid this problem is to use an SSFFS of the pin-bed type, with a matrix of supports, which provides support comparable to a mould-like fixture. The main disadvantages are high cost, and a lack of modularity, which makes them difficult or inefficient to use for parts of differing sizes.

Robotic fixtureless assemblies (RFAs) replace traditional fixtures by robot manipulators equipped with grippers that can cooperatively hold the workpiece [7], [8]. Using RAFs different parts can be manufactured within one work-cell and transitions to other workpieces can be done relatively quickly. However, RAFs have their drawbacks such as high complexity, limited number of robots (and thus holding grasps), and high dependence on software.

The concept described in this papers merges the advantages of RFAs with those of MFFSs, namely: ability to distribute the support action, adaptability to part shapes in a larger range, and high stiffness of the provided support. In our case each fixture element referred to as a physical agent is composed of a mobile robot base, a parallel kinematic machine (PKM) fixed to the mobile platform, an adaptable head with phase-change fluid and an adhesion arrangement, to sustain the supported part perfectly adapting to the part local geometry. The mobility of each support agent and the possibility for the agents to group in regions where some manufacturing operation is being executed results in higher flexibility with lower number of support agents.

Proper fixture design is crucial to product quality in terms of precision, accuracy, and surface finish of the machined parts. Therefore, the research devoted to fixture optimization is quite extensive [2], [9], [10]. Various techniques have been proposed for optimization of fixture layout by formulating different objective functions to determine the location of fixturing supports. In the research for compliant sheet metal parts, Menassa and De Vries [2] use a finite element model of the workpiece to model the deformation, and determine fixture locations by optimizing an objective that is a function of the deformations at the nodes. The design variables are three fixture locators on primary datum as required by the 3-2-1 principle. In [11] an optimization algorithm to obtain the optimal number and location of clamps that minimize the deformation of compliant parts is proposed. Cai *et al*. [9] propose the N-2-1 fixture layout principle for constraining compliant sheet metal parts. This is used instead of the conventional 3-2-1 principle to reduce deformation of sheet-metal parts. They present algorithms for finding the best N locating points such that total deformation of a sheet metal is minimized. They use a finite element model of the part with quadratic interpolation, constraining nodes in contact with the primary datum to only in-plane motion. Nonlinear programming is utilized to obtain the optimal fixture layout. DeMeter [10] introduces a fast support layout optimization model to minimize the maximum displacement-to-tolerance ratio of a set of part features subject to a system of machining loads. The speed-up of the optimization is obtained by a reduced stiffness matrix approach. Most of the previous research related to fixture modeling and design considers fixture in static conditions.

In this paper we propose a concept of the planning system for self adaptable, reconfigurable fixtures composed of mobile support agents. The main role of the planner is to generate the admissible plan of relocation of the mobile agents. It has to find the optimal locations for the supporting heads and the trajectories of the mobile bases that provide continuous support in close proximity to the tool and very high speeds during the relocation phases. In this paper a constrained optimization problem is formulated to find the optimal locations for heads that minimize the given objective function. The constraints to this optimization problem are geometric in nature. The size and dimension of the supporting head are taken into account.

The rest of the paper is organized in the following manner. In Section 2 the concept of a self adaptable reconfigurable fixture system is presented. Section 3 describes an admissible head placement planning problem. In Section 4 head location placement problem is formulated as a constrained nonlinear optimization problem. A numerical example is presented in Section 5. In Section 6 some concluding remarks are presented.

2. Self-Adaptable Reconfigurable Fixture System

Flexible fixture system is composed of mobile robotic agents that can freely move on a bench and reposition below the supported part as shown in Fig. 1. It is assumed that the workpiece is held in position by a subset of locators (not shown in this figure) that remain largely static during the cycle. The remaining agents are highly mobile and change locations to provide additional support in areas affected by the machining process. As mentioned before each support robot consists of a mobile base, a PKM, and an adaptable head. Two mobile agents alternatively supporting a thin sheet while a machine tool with a milling cutter is contouring the workpiece. To simplify motion planning and collision avoidance we assume that the robots move along parallel trajectories. Heads adapt to the local geometry of the workpiece to support it at every repositioning. Adaptation is at two levels: head rotation, to match the approximate orientation of the part surface normal, and head surface deformation, to match the local part surface geometry.

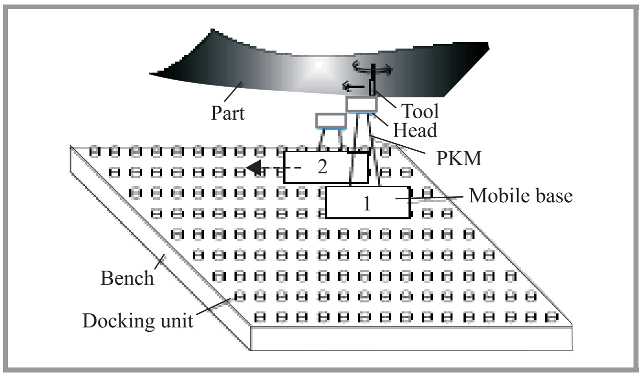


Fig. 1. Self-adaptable reconfigurable fixture system.

The overall goal is to develop the planner, which on the basis of CAD geometric data about the workpiece, representing its state before and after machining, will generate the plan of relocation of the mobile bases and the manip-

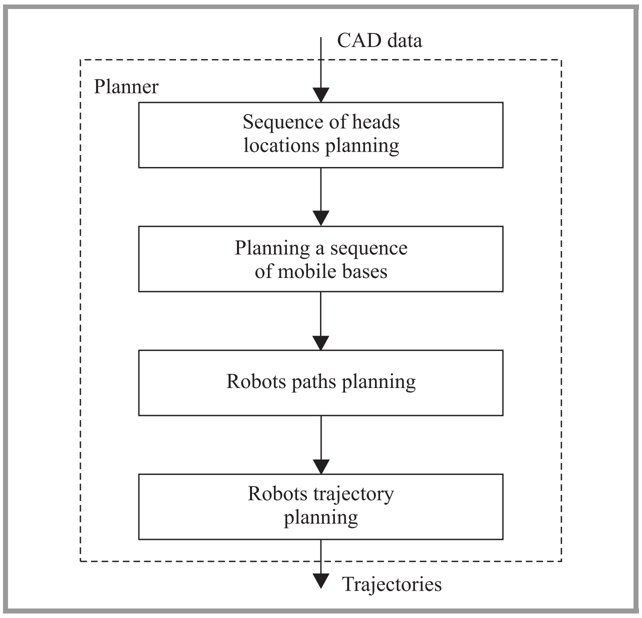

CAD data
Planner
Sequence of heads locations planning
Planning a sequence of mobile bases
Robots paths planning
Robots trajectory planning
Trajectories

Fig. 2. Planner decomposition.

ulators. Planning process is decomposed into four phases: planning a sequence of feasible head placements, planning a corresponding sequence of mobile platforms locations, path planning for mobile platforms and PKMs, trajectory planning for mobile platforms and PKMs (Fig. 2). Obtaining a feasible sequence of head locations is the most difficult part of the planning process. In the paper we will present an approach to solve this problem.

3. Feasible Head Placement

3.1. Geometric Description

We assume that the workpiece contour is modeled as a two-dimensional (2D) simple closed polygonal chain with a given number of linear segments. Closed polygonal curve P in 2D space is described as the ordered set of vertices:

$$P = \{p_1, \dots, p_{M+1}\} = \{(x_1, y_1), \dots, (x_{M+1}, y_{M+1})\}, \quad (1)$$

where the last vertex coincides with the first one, i.e., $p_{M+1} = p_1$. The workpiece boundary consists of M line segments. Each line segment can be described by the following equation:

$$y = a_j x + b_j, \quad j = 1, \dots, M. \quad (2)$$

The coefficients a_j and b_j of the line are calculated from the coordinates of the end points p_j and p_{j+1}:

$$\begin{aligned} a_j &= \frac{y_{j+1} - y_j}{x_{j+1} - x_j}, \\ b_j &= y_j - a_j x_j. \end{aligned} \quad (3)$$

Hereafter, we assume that both heads are identical. The head R is an equilateral triangle

$$R_i = \{r_1, \dots, r_4\}, \quad \text{where } r_4 = r_1. \quad (4)$$

Edge length of the triangle is equal to L.

We assume that the head configuration is specified by $q = (x, y, \theta)^T$, where x, y are Cartesian coordinates relative to a fixed reference coordinate frame and θ is the orientation angle. Configuration space (C-space) of the head is $\mathbb{Q} = \mathbb{R}^2 \times S^1$, where S^1 is the unit circle. Moreover, we explicitly represent the normal vectors for each edge of the head and line segment of the part contour. We denote these normal vectors by $n_k^{R_i}$ for the normal to edge k of the head location i and n_j^P for the normal to j line segment of the polygonal curve P. It should be noted, that for the head edges depend on the orientation θ (but do not depend on x, y-coordinates). In Fig. 3 geometric constraints are depicted.

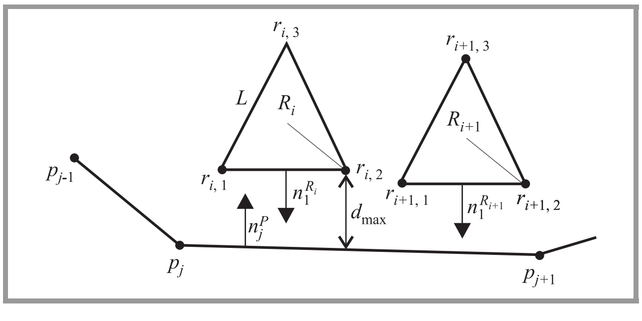


Fig. 3. Geometric constraints for head placement

3.2. Constraints

Four main conditions need to be satisfied for every feasible head placement, R_i:

- The biggest distance between the head and the working profile (workpiece contour) has to be d_{max} to avoid vibrations during contouring.
- The head surface must not come in contact with the tool.
- The maximum allowable distance between two subsequent head locations has to D_{max}.
- The heads must not overlap each other.

To satisfy these conditions we must to know the minimum and maximum distance between two objects. Minimum distance calculation is essential for collision detection, if the the minimum distance between to objects is zero, then they are in contact. The distance between two polytopes (in 2D polygons) Q and P is defined as

$$d_m(P, Q) = \min_{p \in P, q \in Q} \|p - q\|. \quad (5)$$

Expression (5) can be reformulated in terms of the Minkowski difference of two polytopes, i.e.,

$$P \ominus Q = \{z | z = p - q, p \in P, q \in Q\} = Z. \quad (6)$$

Using Eq. (6) we can rewrite Eq. (5) as

$$d_m(P, Q) = \min_{p \in P, q \in Q} \|p - q\| = \min_{z \in P \ominus Q} \|z\| \quad (7)$$

and we have reduced the problem of computing distance between two polytopes to the problem of computing the minimum distance from one polytope to the origin of the coordinate frame. The Minkowski difference of two convex polytopes is itself convex polytope. Since $Z = P \ominus Q$ is a convex set, and since the norm, $\|z\|$, is a convex function, $\hat{z} = \arg\min_{z \in Z} \|z\|$ is unique. However, p and q to achieve this minimum are not necessarily unique. To compute the minimum distance we use well-known GJK algorithm [12]. The Euclidean distance d from point $p_k = (x_k, y_k)^T$ to the line segment $y = a_j x + b_j$ can be calculated by the following expression:

$$d = \frac{|y_k - a_j x - b_j|}{\sqrt{1 + a_j^2}}. \quad (8)$$

The biggest allowable distance between the head and the working profile has to be d_{max} to avoid vibrations during contouring

$$d_i(P,R_i) \leqslant d_{max}, \quad i, = 1,\dots,N-1 \tag{9}$$

This means that the distance between workpiece contour and the closest edge $E_k^{R_i}$ of the head R_i to the contour segment must not be greater than d_{max}. The heads must not overlap each other

$$\mathrm{int}(R_i) \cap \mathrm{int}(R_{i+1}) = \emptyset, \quad i, = 1,\dots,N-1\,, \tag{10}$$

where $\mathrm{int}(R_i)$ denotes the interior of the triangle. However, two heads may contact each other. Contact between two heads can occur only when orientation θ satisfies the following condition

$$\begin{gathered}(r_{i,j-1}(\theta_i) - r_{i,j}(\theta_i)) \cdot n_k^{R_{i+1}}(\theta_{i+1}) \geqslant 0 \wedge \\ (r_{i,j+1}(\theta_i) - r_{i,j}(\theta_i)) \cdot n_k^{R_{i+1}}(\theta_{i+1}) \geqslant 0, \\ j,k = 1,2,3;\ i, = 1,\dots,N-1\,.\end{gathered} \tag{11}$$

If this condition satisfied there is a contact between edge, $E_k^{R_{i+1}}$, of the head R_{i+1} and vertex $r_{i,j}$ of the head R_i. At extreme, the vertices $r_{i,j}$ and $r_{i+1,k}$ coincide, while at the other extreme, vertices $r_{i,j}$ and $r_{i+1,k+1}$ coincide.
Analogously, when the condition

$$\begin{gathered}(r_{i+1,j-1}(\theta_{i+1}) - r_{i+1,j}(\theta_{i+1})) \cdot n_k^{R_i}(\theta_i) \geqslant 0 \wedge \\ (r_{i+1,j+1}(\theta_{i+1}) - r_{i+1,j}(\theta_{i+1})) \cdot n_k^{R_i}(\theta_i) \geqslant 0, \\ j,k = 1,2,3;\ i, = 1,\dots,N-1\end{gathered} \tag{12}$$

is satisfied there is a contact between edge, $E_k^{R_i}$, of the head R_i and vertex $r_{i+1,j}$ of the head R_{i+1}. Again, at extreme, the vertices $r_{i+1,j}$ and $r_{i,k}$ coincide, while at the other extreme, vertices $r_{i+1,j}$ and $r_{i,k+1}$ coincide. The head surface must not come in contact with the tool

$$d_i(P,R_i) \geqslant d_{min}, \quad i = 1,\dots,N\,. \tag{13}$$

4. An Optimization Problem

Planning a sequence of the supporting heads locations can be formulated as a constrained optimization problem. The optimization model is presented as follows:

- Design variables. The head locations $R_i(x,y,\theta)$, $(i = 1,\dots,N)$. Hence, the vector of variables is defined as follows

$$\boldsymbol{x} = [x_1, y_1, \theta_1, \dots, x_N, y_N, \theta_N]^T\,.$$

- Min-max nonlinear optimization problem:

$$\min \max f_i(\boldsymbol{x}), \quad i = 1,\dots,N, \tag{14}$$

where $f_i(x) = d_i^2(P,R_i)$ is the squared distance of the head $R_i, i = 1,\dots,N$ to the contour P. The following motivation is behind this form the objective function: the closest distance of the support head to the working contour the lowest vibrations may occur.

- Constraints. All previously defined constraints can be described in general form

$$\boldsymbol{g}(\boldsymbol{x}) \leqslant 0\,. \tag{15}$$

Moreover, the following linear inequality constraints must be satisfied

$$A\boldsymbol{x} - b \leqslant 0\,, \tag{16}$$

where the entries of the matrix A and the vector b are calculated according to Eq. (3). It means that the heads in each location must be inside the region limited by the working contour.

To solve the nonlinear min-max optimization problem Eqs. (14)–(16) in an efficient and robust way we transform this problem into a special nonlinear programming problem (NLP). We introduce one additional variable, z, and N additional nonlinear inequality constraints in the form

$$f_i(\boldsymbol{x}) - z \leqslant 0, \quad i = 1,\dots,N\,. \tag{17}$$

The following equivalent optimization problem can be defined

$$\min z \tag{18}$$

subject to the constraints of the original problem Eqs. (14)–(16) and the additional constraints (17). To solve this problem an efficient existing nonlinear programming techniques can be used.

5. A Numerical Example

Let us consider a workpiece which boundary is shown in Fig. 4. This contour can be described as a closed polygonal chain. The vertices are enumerated in anticlockwise direction and their Cartesian coordinates are given in Table 1. The following values of the parameters are selected: edge length of the head $L = 70$ mm, maximum distance $d_{\max} = 20$ mm, minimum distance $d_{\min} = 1$ mm of the head to the workpiece contour, and maximum distance between two heads $D_{\max} = 20$ mm. The number of

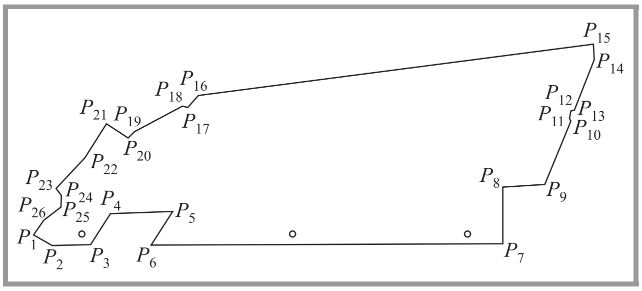


Fig. 4. Workpiece boundary.

variables in this specific problem is equal to $N = 268$ that corresponds to 89 head locations. The number of nonlinear inequality constraints is 882 and linear inequality constraints is 26. The code solving an optimization problem was implemented in Matlab. The specific optimization algorithm used is the constrained nonlinear programming function *fmincon()* from Matlab [13]. The main problem is to find a feasible starting point for the optimization algorithm, which satisfies all constraints. The choice of the starting point strongly influence the performance. Typically, to solve this optimization problem 20-25 iterations are required. The value of termination tolerance is equal to $1 \cdot 10^{-6}$. The preliminary optimization results are shown in Fig. 5. This figure presents the admissible head placement obtained by solving NLP problem.

Table 1
Vertices of the contour

Point	x [cm]	y [cm]	Point	x [cm]	y [cm]
P_1	8.30	75.54	P_{14}	283.96	162.36
P_2	16.51	70.28	P_{15}	284.01	169.15
P_3	36.07	70.33	P_{16}	88.67	143.77
P_4	45.83	85.86	P_{17}	83.90	138.10
P_5	76.29	87.22	P_{18}	81.54	138.94
P_6	65.50	70.42	P_{19}	57.26	126.08
P_7	239.31	70.92	P_{20}	54.52	123.28
P_8	239.31	99.13	P_{21}	43.82	130.17
P_9	259.93	100.08	P_{22}	32.99	113.25
P_{10}	272.51	131.75	P_{23}	19.14	98.38
P_{11}	271.78	132.71	P_{24}	21.29	94.92
P_{12}	273.09	136.31	P_{25}	21.24	89.02
P_{13}	274.49	136.97	P_{26}	12.93	82.78

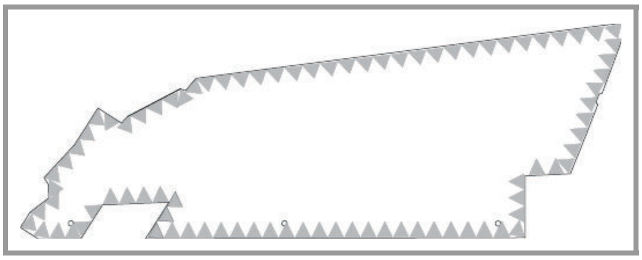


Fig. 5. An admissible head placement.

6. Summary and Conclusion

In this paper, we presented a methodology for modeling and optimization for self adaptable, reconfigurable fixtures supporting thin sheet metal parts to minimize part dimensional deformation during milling. Compliant sheet metal parts are widely used in various manufacturing processes including automotive and aerospace industries. The concept of the multi-layer planning system is proposed. The most difficult part of the planning process, namely, a head placement problem is considered. To find a feasible plan of a sequence of supporting head locations nonlinear programming problem is solved. Finally, a numerical example is used to illustrate the feasibility of this method. In future work, we will develop a complete planner including trajectory planning of the mobile bases and PKMs.

Acknowledgements

This work was partially supported by the Polish Ministry of Science and Higher Education through grant No. N N 514 1287 33.

References

[1] A. Vidal, M. Alberti, J. Ciurana, and M. Casadesús, "A decision support system for optimising the selection of parameters when planning milling operations", *Int. J. Mach. Tool Manu.*, vol. 45, no. 2, pp. 201–210, 2005.

[2] R. Menassa and W. De Vries, "Optimization methods applied to selecting support positions in fixture design", *ASME J. Eng. Ind.*, vol. 113, no. 4, pp. 412–418, 1991.

[3] S. Vallapuzha, E. C. De Meter, S. Choudhuri, and R. P. Khetan, "An investigation of the effectiveness of fixture layout optimization methods" *Int. J. Mach Tool Manu.*, vol. 42, no 2, pp. 251–263, 2002.

[4] M. N. Sela, O. Gaudry, E. Dombre, and B. Benhabib, "A reconfigurable modular fixturing system for thin-walled flexible objects" *Int. J. Adv. Manuf. Technol.*, vol. 13, no. 9, pp. 611–617, 1997.

[5] B. Shirinzadeh and Y. Tie, "Experimental investigation of the performance of a reconfigurable fixture system", *Int. J. Adv. Manuf. Technol.*, vol. 10, no. 5, pp. 330–341, 1995.

[6] K. Youcef-Toumi, W. Liu, and H. Asada, "Computer-aided analysis of reconfigurable fixtures and sheet metal parts for robotics drilling", *Rob. Comp.-Integr. Manu.*, vol. 4, pp. 3–4, pp. 387–393, 1988.

[7] Z. M. Bi and W. J. Zhang, "Flexible fixture design and automation: review, issues and future directions", *Int. J. Prod. Res.*, vol. 39, no. 13, pp. 2867–2894, 2001.

[8] Y. Kang, Y. Rong, J. Yang, and W. Ma, "Computer-aided fixture design verification", *Assembly Autom.*, vol. 22, no. 4, pp. 350–359, 2002.

[9] W. Cai, S. J. Hu, and J. X. Yuan, "Deformable sheet metal fixturing: principles, algorithms, and simulations", *T ASME, J. Manuf. Sci. Eng.*, vol. 118, no. 3, pp. 318–324, 1996.

[10] E. C. DeMeter, "Fast support layout optimization", *Int. J. Mach. Tools Manuf.*, vol.38, no. 10–11, pp. 1221–1239, 1998.

[11] M. R. Rearick, S. J. Hu, and S. M. Wu, "Optimal fixture design for deformable sheet metal workpieces", *Trans. NAMRI/SME*, vol. 21, pp. 407–412, 1993.

[12] E. Gilbert, D. Johnson, and S. Keerthi, "A fast procedure for computing the distance between objects in three-dimensional space", *IEEE J. Robot. Autom.*, vol. 4, pp. 193–203, 1988.

[13] Matlab. *Optimization Toolbox User's Guide*. The MathWorks, Inc., 2009.

Wojciech Szynkiewicz received the Ph.D. degree in robotics in 1996 from the Warsaw University of Technology (WUT). He is an Assistant Professor employed by the Institute of Control and Computation Engineering of WUT. From 1999 to 2003 he was the Deputy Director and Secretary to the Scientific Council of the Research Center for Automation and Information-Decision Technology-CATID. His research activities concentrate on multi-robot/multi-agent systems, motion planning, autonomous mobile robots, robot controller structures, and real-time and distributed systems. He works on sensor-based motion planning and control algorithms for multi-robot systems, including service, personal and mobile robots.
email: W.Szynkiewicz@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00–665 Warsaw, Poland

Paper

Performance Analysis of Hybrid Phase Shift Keying over Generalized Nakagami Fading Channels

Mahmoud Youssuf and Mohamed Z. Abdelmageed

Abstract—In addition to the benefits of hybrid phase shift keying (HPSK) modulation in reducing the peak to average power ratio of the transmitted signal to reduce the zero crossings and the 0°-degree phase transmissions, HPSK enhances the bit error rate (BER) measure of the signal performance. The constellation of the HPSK is analyzed, and an expression for the conditional probability of HPSK modulation over additive white Gaussian noise (AWGN) is derived. This BER measure of HPSK is shown to outperform quadrature phase shift keying (QPSK) modulation. HPSK performance through Nakagami – m fading channel is also considered.

Keywords—bit error rate, hybrid phase shift keying, quadrature phase shift keying.

1. Introduction

Hybrid phase shift keying (HPSK) is used in wideband code division multiple access (WCDMA) systems thanks to its low peak-to-average power ratio. This low ratio of peak-to-average power results in reducing the number of zero crossing of phase transitions of the output transmitted signal. In this paper we prove that HPSK outperforms other quadrature modulation techniques, such as offset quadrature phase shift keying (OQPSK) by 3 dB.

The paper is organized as follows: Section 2 describes the HPSK constellation in case of two channels at different amplitudes. Section 3 derives an expression for the conditional probability of error of HPSK modulated signal over an additive white Gaussian noise (AWGN). In Section 4 we apply the obtained expression in evaluating the performance of HPSK modulated signal over a generalized Nakagami – m channel. Finally, Section 5 includes numerical results and comments.

2. The Hybrid Phase Shift Keying Constellation

The HPSK has been proposed as the spreading technique for WCDMA to eliminate the zero crossings for every other signal transition and to eliminate the 0°-degree phase shift transitions for every other chip point, as well as to improve the bit error rate (BER) measure of the direct sequence WCDMA (DS-WCDMA) system performance. In 3G systems the mobile station (MS) can transmit more than one channel. The different channels are assigned to either I or Q path.

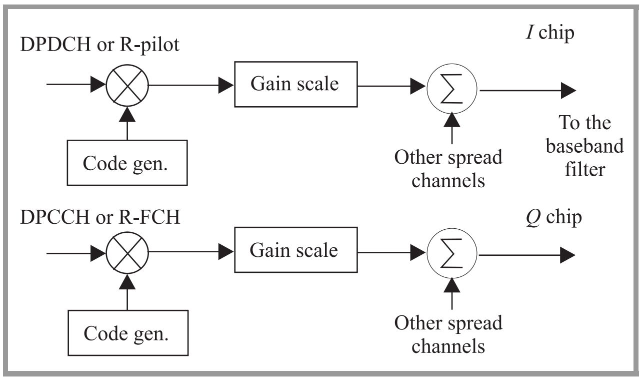


Fig. 1. The basic reverse channel structure of 3G system. Explanations: DPCCH – dedicated physical control channel, DPDCH – dedicated physical data channel, R-FCH – reverse fundamental channel.

In the case of transmitting only two channels, as in Fig. 1, one of the channels (DPDCH or R-pilot) is applied to the I path and the other channel (DPCCH or R-FCH) is applied to the Q path [1]. Additional, high data rate channels are combined alternatively on the I or Q paths. Each channel is spread by a different orthogonal even-numbered Walsh code. In the general case the channels can be at different power levels, as in Fig. 2 which maps onto a rectangular four quadrature amplitude modulation (4-QAM) constellation.

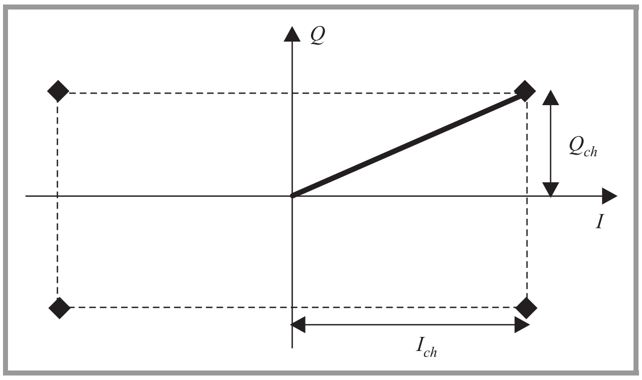


Fig. 2. The 4-QAM constellation for two channels at different amplitudes.

The QAM signal waveforms may be expressed as

$$\begin{aligned} S_m(t) &= \mathrm{Re}\left\{A e^{j\theta_m} g(t) e^{j2\pi f_c t}\right\} \\ &= A_{ch} g(t)\left[\cos(2\pi f_c t)\cos(\theta_m)\right. \\ &\quad \left. -\sin(2\pi f_c t)\sin(\theta_m)\right], \end{aligned} \tag{1}$$

where: $A_{ch} = \sqrt{I_{ch}^2 + Q_{ch}^2}$, where I_{ch} and Q_{ch} are the information bearing signals amplitudes of the I path channel and Q path channel, respectively, $g(t)$ is the pulse shape signal, $\theta_m = \tan^{-1}\frac{Q_{ch}}{I_{ch}}$ is the phase of the signal vector and it varies with $m = 1, 2, 3, 4$.

So, the QAM signal waveform may be viewed as a linear combination of two orthogonal wave forms $f_1(t)$ and $f_2(t)$ such that [2]:

$$S_m(t) = S_{m1} f_1(t) + S_{m2} f_2(t), \tag{2}$$

where S_{m1} and S_{m2} are the component of the two dimensional vector S_m:

$$S_m = \left[S_{m1} \quad S_{m2}\right] = \left[I_{ch} \quad Q_{ch}\right]. \tag{3}$$

In the 4-QAM according to the position of the vector point and according to the value of I_{ch} and Q_{ch}:

$$\therefore S_m = \left[A_{ch}\cos\theta_m \quad A_{ch}\sin\theta_m\right]. \tag{4}$$

In the reverse link of DS-CDMA systems the I_{ch} and Q_{ch} are complex scrambled with a complex scrambling signal $(I_s + jQ_s)$ as in Fig. 3.

The final I and Q signals are produced mathematically by the multiplication of the two complex signals; the complex data signal $(I_D + jQ_D)$ which has already spread into chips $(I_{ch} + jQ_{ch})$, and the complex scrambling signal $(I_s + jQ_s)$ so the final I and Q signals are:

$$I + jQ = (I_{ch}I_s - Q_{ch}Q_s) + j(I_{ch}Q_s + Q_{ch}I_s) = A_{ch}A_s e^{j(\phi_{ch}+\phi_s)},$$

$$I = A_{ch}A_s\cos\left(\frac{\pi}{M}(2n-1) + \theta_m\right) = A\cos\left(\frac{\pi}{M}(2n-1) + \theta_m\right), \tag{5}$$

$$Q = A_{ch}A_s\sin\left(\frac{\pi}{M}(2n-1) + \theta_m\right) = A\sin\left(\frac{\pi}{M}(2n-1) + \theta_m\right). \tag{6}$$

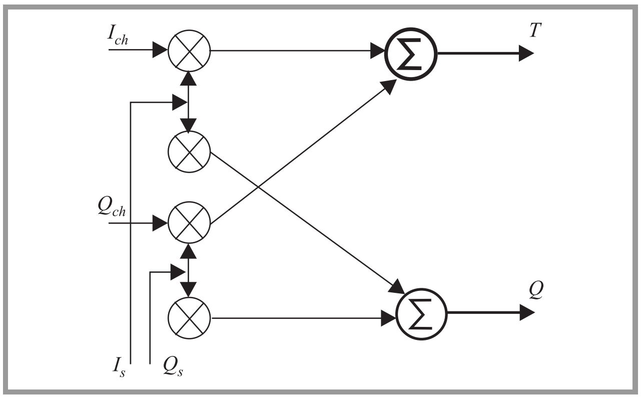


Fig. 3. Complex scrambling of HPSK.

Since final constellation is formed by complex multiplication of the two signals of chip constellation and scrambling constellation which is always QPSK constellation as in Fig. 4, then: from Eqs. (5) and (6) we conclude that the final signal $(I + jQ)$ is a QPSK constellation with two dimensional vector representation S_{mn}, where:

$$S_{mn} = \left[A\cos\left(\frac{\pi}{M}(2n-1) + \theta_m\right) \quad A\sin\left(\frac{\pi}{M}(2n-1) + \theta_m\right)\right], \tag{7}$$

where $A = A_{ch}A_s$; $n = 1, 2, 3, 4$; $m = 1, 2, 3, 4$ and $M = 4$.

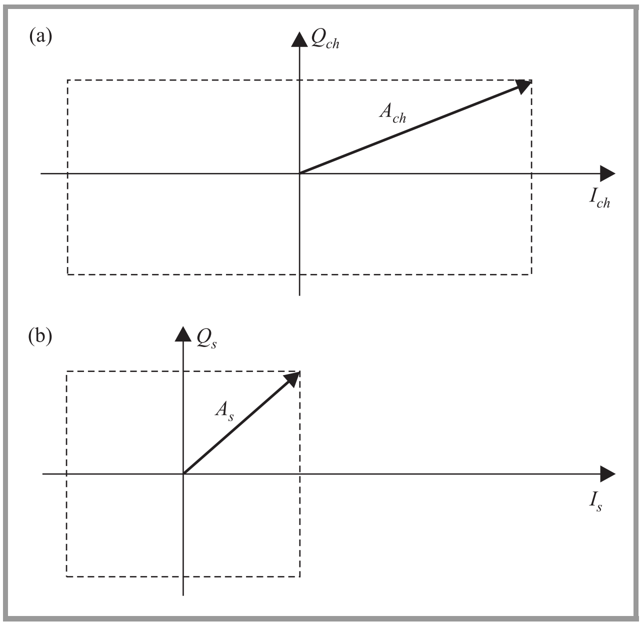


Fig. 4. The chip constellation (a) and the scrambling constellation (b).

This new constellation has points that rotate according to the angle: $\frac{\pi}{4}(2n-1) + \theta_m$, while θ_m changes according to the value of I_{ch} and Q_{ch}, for example, if $n = 1$, the point (I_{ch}, Q_{ch}) will be transferred by the angle equal to $(45^\circ + \theta_m)$. So the new constellation is really an eight point constellation with two independent QPSK constellation as shown in Fig. 5 according to the value of θ_m. One of these two constellations corresponds to $\theta_m > 45^\circ$ and it rotates by angle equal to $\phi_1 = (\theta_m - 45^\circ)$ from the original axes. The other QPSK constellation corresponds to $\theta_m < 45^\circ$ and rotates with angle equal to $\phi_2 = -(\theta_m - 45^\circ)$ from the original axes. So the new final constellation consists of two independent QPSK constellations with complex scrambling:

$$S_n = \left[A\cos\left[\frac{2\pi}{M}(n-1) + \phi_1\right] \quad A\sin\left[\frac{2\pi}{M}(n-1) + \phi_1\right]\right], \tag{8}$$

$$S'_n = \left[A\cos\left[\frac{2\pi}{M}(n-1) + \phi_2\right] \quad A\sin\left[\frac{2\pi}{M}(n-1) + \phi_2\right]\right]. \tag{9}$$

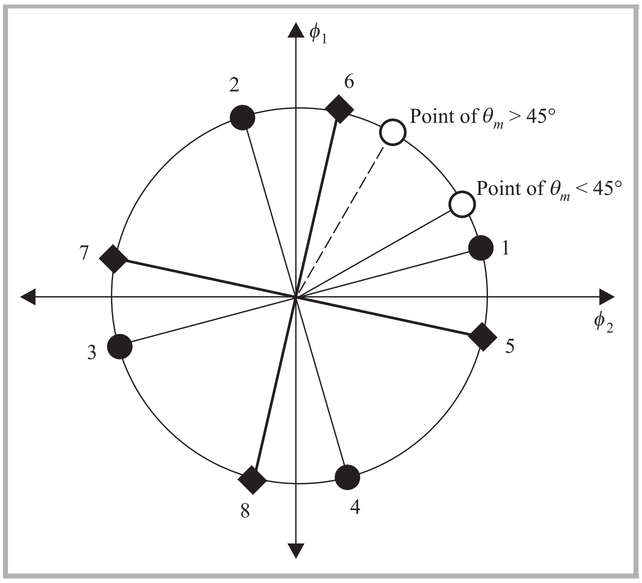


Fig. 5. The final constellation of the scrambled chip of different channel amplitudes.

So, the final constellation has eight points, as in Fig. 5, with the angular distribution determined by the relative levels of the two channels signals.

3. Probability of Error of Hybrid Phase Shift Keying Modulated Signal over Additive White Gaussian Noise

We concluded in the previous section that the final constellation of the two channels at different amplitudes consists of two independent QPSK constellations with complex scrambling. The average value of the amplitude of the new constellation is $\sqrt{2}$ times the value of the amplitude of the traditional QPSK. To obtain the BER of the HPSK modulated signal as a measure of its performance we will assume that this signal is transmitted over AWGN channel. The received signal is demodulated with correlated demodulator or a matched filter demodulator and the decision for the received observation vector $r = [r_1, r_2, r_3, \ldots, r_N]$ among all the transmitted signals S_m is based on the maximum of the conditional probability distribution function (pdf) $P(r/S_m)$, which is the maximum likelihood (ML).

The optimum ML detector computes a set of M correlation metrics $C(r, S_m) = -2rS_m$ and selects the signal corresponding to the largest correlation metric [3].

Applying this metric in our study case, r is the received signal vector $r = [r_1 \;\; r_2]$ which is projected onto each of the four possible transmitted signals vectors S_m for $m = 1,\, 2,\, 3,\, 4,$ where $M = 4$. We can consider that the correlation detector in the case of HPSK modulated signal is equivalent to a phase detector that computes the phase of the received signal from r and selects the signal S_m whose phase is closest to r, where the phase of r is $\theta_r = \tan^{-1}\frac{r_2}{r_1}$.

The probability of error can be computed if we determine the power density function of θ_r $P_{\theta_r}(r)$. We consider the case in which the transmitted signal phase is equal to zero, as in Fig. 6.

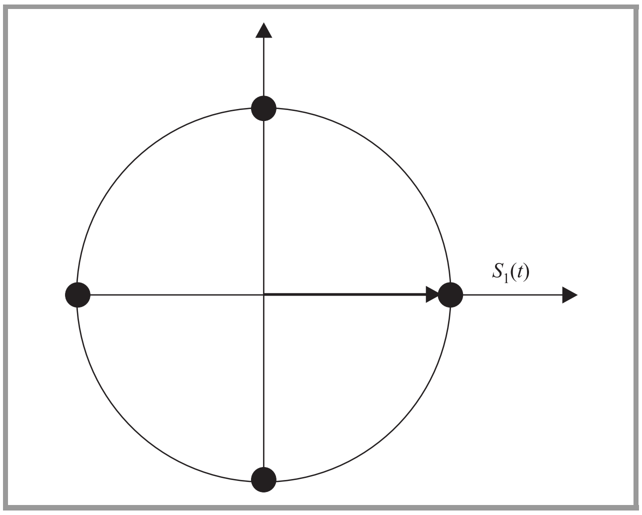


Fig. 6. The HPSK vector constellation.

The transmitted vector $S_1 = \left[\sqrt{\varepsilon_s} \;\; 0\right]$, where ε_s is the energy of the transmitted HPSK signal $S_1(t)$. The received signal vector is $r = [r_1 \;\; r_2] = \left[\sqrt{\varepsilon_s} + n_1 \;\; n_2\right]$, where n_1 and n_2 are jointly Gaussian random variables with mean and variances $E(r_1) = \sqrt{\varepsilon_s},\ E(r_2) = 0$ and $\sigma r_{12} = \sigma r_{22} = \sigma r_2 = \frac{1}{2}N_0$. Consequently, the joint pdf of r_1 and r_2 is:

$$P(r_1, r_2) = \left(\frac{1}{\pi N_0}\right) \exp\left[-\frac{(r_1 - \sqrt{\varepsilon_s})^2 + r_2^2}{N_0}\right]. \tag{10}$$

The pdf of the phase θ_r is obtained by a change in variables from (r_1, r_2) to $(A,\, \theta_r)$, where:

$$A = \sqrt{r_1^2 + r_2^2}, \quad \theta_r = \tan^{-1}\frac{r_2}{r_1},$$

$$P(A,\, \theta_r) = A\left(\frac{1}{\pi N_0}\right) \exp\left[-\frac{(A^2 + \varepsilon_s - 2\sqrt{\varepsilon_s} A \cos\theta_r)}{N_0}\right], \tag{11}$$

$$P_{\theta_r}(\theta_r) = \int_0^\infty P(A,\, \theta_r)\mathrm{d}A = \frac{1}{2\pi}\mathrm{e}^{-2\gamma_s \sin^2\theta} \int_0^\infty A\mathrm{e}^{-\frac{A - \sqrt{4\gamma_s}\cos\theta)^2}{2}}, \tag{12}$$

where $\gamma_s = \frac{\varepsilon_s}{N_0}$ is the signal-to-noise ratio (SNR).

For large values of $\gamma_s \gg 1$ and $|\theta_r| \leq \frac{\pi}{2}$, $P_{\theta_r}(\theta_r)$ is well approximated as

$$P_{\theta_r}(\theta_r) = \frac{1}{2\pi} e^{-2\gamma_s \sin^2\theta_r} I\,, \tag{13}$$

where

$$I = \int\limits_0^\infty A \exp\left(-\frac{(A - \cos\theta_r\sqrt{4\gamma_s})^2}{2}\right) \mathrm{d}A\,, \tag{14}$$

$$I = \cos\theta_r \sqrt{2\pi}\sqrt{4\gamma_s}\,, \tag{15}$$

$$P_{\theta_r}(\theta_r) = \sqrt{\frac{2\gamma_s}{\pi}} \cos\theta_r \mathrm{e}^{-2\gamma_s \sin^2\theta_r}\,. \tag{16}$$

When $s_1(t)$ is transmitted a decision of error is made if the noise causes the phase to fall outside the range $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$, and hence the probability of a symbol error is:

$$P_4 = 1 - \int\limits_{-\frac{\pi}{4}}^{\frac{\pi}{4}} P_{\theta_r}(\theta_r)\mathrm{d}\theta_r\,. \tag{17}$$

By substituting for $P_{\theta_r}(\theta_r)$ and performing the change of variables from θ_r to μ, where $\mu = \sin\theta_r\sqrt{2\gamma_s}$, we find

$$P_4 = 2Q\left(\sqrt{2\gamma_s \sin^2\left(\frac{\pi}{4}\right)}\right) = 2Q\left(\sqrt{\gamma_s}\right). \tag{18}$$

In case of HPSK

$$\varepsilon_s = 2(k\varepsilon_b) = 4\varepsilon_b \;\therefore P_4 = 2Q\left(\sqrt{\frac{4\varepsilon_b}{N_0}}\right). \tag{19}$$

Since the transmitted signals represented by the vector points of the final constellation are equally likely to be transmitted and since the 8-points are distributed around a circle consisting of two independent QPSK constellations with complex scrambling then the average probability of the symbol error in the case of two channels at different amplitudes is:

$$P_{symbol} = \frac{1}{2}\left[P_1(e) + P_2(e)\right], \tag{20}$$

where $P_1(e)$ is the probability of symbol error of the first QPSK constellation, $P_2(e)$ is the probability of symbol error of the second constellation:

$$P_S = 2Q\left(\sqrt{\frac{4\varepsilon_b}{N_0}}\right). \tag{21}$$

The bit error probability in this case is:

$$P_b = Q\left(\sqrt{\frac{4\varepsilon_b}{N_0}}\right) = Q\left(\sqrt{4\gamma}\right), \tag{22}$$

where γ is the SNR $= \varepsilon_b/N_0$.

It is simply interesting to compare the performance of HPSK with that of QPSK since both types of signals are two dimensional. Since the error probability is dominant by the arguments of the Q function, we may simply compare the arguments of Q for the two signal formats of HPSK and QPSK. The ratio of the two arguments is equal $R = 2$. So, HPSK has SNR advantage of 10 log 2 ≈ (3 dB) over QPSK.

4. The Performance of HPSK Signal over a Generalized Nakagami – m Channel

The mobile communication channel is noisy, multipath and is subjected to fading. The channel fading conditions depend on the propagation conditions and the clutter types the waves propagate through. In some cases the fading can be more severe than Rayleigh, while in other cases where line of sight or near line of sight conditions is available, the signal will be more stable. The Nakagami distribution is shown to fit results more generally than other distributions [4]. In this section we will evaluate the average BER of HPSK systems subjected to Nakagami fading. We will evaluate the expected value of the conditional P_b as given by Eq. (22) over Nakagami distribution. In Nakagami channel the path amplitude probability density function is given by

$$f_r(r) = \frac{2}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m r^{2m-1} \exp\left(-\frac{m}{\Omega}r^2\right), \tag{23}$$

where m is the fading parameter and it describes the fading severity which is defined as the ratio of moments or it is the ratio of the square of the mean signal power to the variance of the signal power

$$m = \frac{\Omega^2}{E\left[r^2 - \Omega\right]}, \quad m \geq 0.5, \quad \Omega = E\left[r^2\right]. \tag{24}$$

The received signal power γ follows gamma distribution and its pdf is given by

$$f_r(\gamma) = \left(\frac{m}{\Omega}\right)^m \frac{\gamma^{\,m-1}}{\Gamma(m)} \exp\left(-\frac{m}{\Omega}\gamma\right); \quad \gamma \geq 0,\; m \geq 0.5\,.$$

The average probability of error will be given by

$$P(e) = \int\limits_0^\infty P(e/\gamma) f_\gamma(\gamma)\mathrm{d}\gamma = \int\limits_0^\infty Q\left(\sqrt{4\gamma}\right) f_\gamma(\gamma)\mathrm{d}\gamma. \tag{25}$$

The integral of the average probability is evaluated in [5]. The average probability of error is finally given by

$$P_e = \left(\frac{1}{2\sqrt{\pi}}\right) \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m+1)} \left(\frac{2m}{2m+4\Omega}\right)^m \times {}_2F_1\left(\frac{1}{2}, m;\; m+1;\; \frac{2m}{2m+4\Omega}\right), \tag{26}$$

where ${}_2F_1(a, b;\; c, x)$ is the Gauss hypergeometric function [6] defined by

$${}_2F_1(a, b;\; c, x) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{x^n}{\Gamma(n+1)} \tag{27}$$

or with integral representation:

$${}_2F_1(a, b;\; c, x) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int\limits_0^1 t^{b-1}(1-t)^{c-b-1}(1-xt)^{-a}\mathrm{d}t\,, \tag{28}$$

where $c > b > 0$.

The integral representation of ${}_2F_1$ is valid under the assumption that $|x| \leq 1$. In our case, it reduces to:

$$ {}_2F_1\left(\frac{1}{2}, m; m+1; x\right) = \frac{\Gamma(m+1)}{\Gamma(m)} \int\limits_0^1 \frac{t^{m-1}}{\sqrt{(1-xt)}}\,\mathrm{d}t\,. \qquad (29) $$

The average BER can then be given by

$$ P_e = \left(\frac{1}{2\sqrt{\pi}}\right) \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m)}\, x^m \int\limits_0^1 \frac{t^{m-1}}{\sqrt{(1-xt)}}\,\mathrm{d}t\,, \qquad (30) $$

where

$$ x = \frac{2m}{2m+4\Omega}. \qquad (31) $$

The substitution $(t = \sin 2(\theta)/x)$, is useful to solve the integral. Finally, we have for arbitrary value of m:

$$ P_e = \left(\frac{1}{\sqrt{\pi}}\right) \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m)} \int\limits_0^{\arcsin(\sqrt{x})} \sin^{2m-1}\theta..\mathrm{d}\theta \qquad (32) $$

and for integer m, Eq. (32) will result into:

$$ P_e = \left(\frac{1}{\sqrt{\pi}}\right) \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m)} \sum_{n=0}^{m-1} \binom{m-1}{n} \frac{(-1)^n}{2n+1} \left[1 - y^{(n+\frac{1}{2})}\right], \qquad (33) $$

where

$$ y = \sqrt{1-x} = \sqrt{\frac{4\Omega}{2m+4\Omega}}. \qquad (34) $$

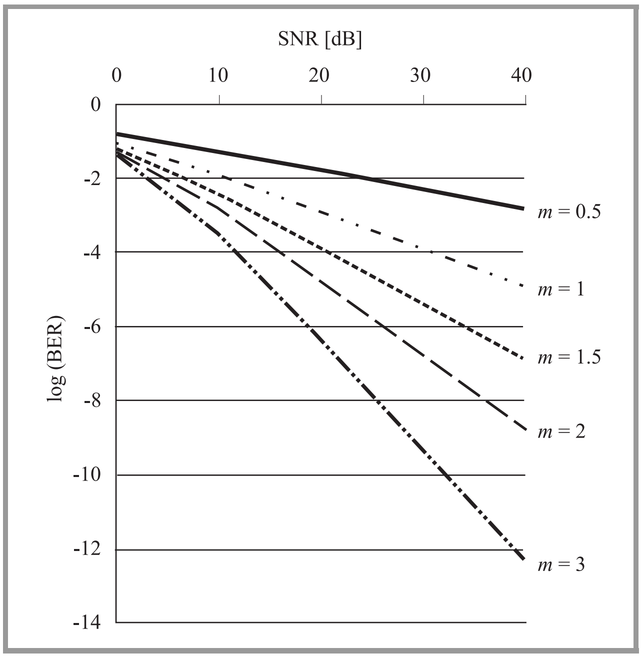


Fig. 7. The BER versus SNR of HPSK in Nakagami fading channels.

Special cases. The probability of error is computed for different value of m (Fig. 7). For the case of severe fading:

- $m = \frac{1}{2}$ (the half Gaussian distribution) Eq. (32) reduces to:
$$ P_e = \arcsin(\sqrt{x})/\pi\,, \qquad (35) $$
- $m = 3/2$, Eq. (33) will reduce to:
$$ P_e = \left(\arcsin(\sqrt{x} - \sqrt{x(1-x)})\right)/\pi\,, \qquad (36) $$
- $m = 1$ (the case of Rayleigh fading), the average probability of error in Eq. (33) reduces to:
$$ P_e = (1-y)/2\,, \qquad (37) $$
- $m = 2$, Eq. (33) will result to:
$$ P_e = (2 - 3y + y^3)/4\,, \qquad (38) $$
- $m = 3$ (very close to Rician fading), Eq. (33) will result to:
$$ P_e = \left(8 - (15y - 10y^3 + 3y^5)\right)/16\,. \qquad (39) $$

5. Conclusion

In this paper we come to an easy to evaluate expressions for the BER of HPSK performance in Nakagami fading channel as shown in the previous section. The cases of more severe fading than Rayleigh ($m = 1$) and Racian ($m = 3$) fading are considered. Figure 7 shows the probability of error of HPSK for different values of m. Also we proved that HPSK has SNR advantage of nearly 3 dB over QPSK.

References

[1] "HPSK spreading for 3G", Application Note. Agilent Technol. Inc., 2003.

[2] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 2001.

[3] Lowell Hoover, "Deriving the Quadratic Modulators", PAN101A & AN102A. Bloomington: Polyphase Microwave Inc., 2005.

[4] U. Charach, "Reception through Nakagami fading multipath channels with randam delays", *IEEE Trans. Commun.*, vol. Com-27, no. 4, pp. 657–670, 1979.

[5] E. K. Al Hussaini and A. Al-Bassioni, "Performance of MRC diversity systems for the detection of signals with Nakagami fading", *IEEE Trans. Commun.*, vol. Com-33, no. 12, pp. 1315–1319, 1985.

[6] L. C. Andrews, *Special Functions for Engineers and Applied Mathematicians*. London: Macmillan Publ., 1985.

Mahmoud Youssuf is a Ph.D. candidate received his Master of science degree in electrical engineering at Alazhar University, Egypt. He is a Lecturer at the Department of Systems and Computers at Alazhar University. His research interests include issues related to wireless communications, federal networks, satellite communication applications. He is an author of research studies published at national and international journals, conference proceedings.
e-mail: mahmoud.youssuf@gmail.com
Alazhar University
Faculty of Engineering
Department of Systems and Computers
Republic Palace Square
Alawkaf Buildings A-1, 4th floor
Hadyek Elqubba, Cairo, Egypt

Mohamed Zaki Abdelmageed received his Ph.D. degree in electrical engineering. He is the chief of the Department of Systems and Computers at Alazhar University, Egypt. His research interests are related to networks, wireless communications and computer systems. He has published research papers at different national and international journals, conference proceedings.
e-mail: azhhar@mailer.eun.eg
Alazhar University
Faculty of Engineering
Department of Systems and Computers
Republic Palace Square
Alawkaf Buildings A-1, 4th floor
Hadyek Elqubba, Cairo, Egypt

Information for Authors

Journal of Telecommunications and Information Technology **(JTIT)** is published quarterly. It comprises original contributions, dealing with a wide range of topics related to telecommunications and information technology. **All papers are subject to peer review**. Topics presented in the JTIT report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

JTIT is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, voice communications devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology.

Suitable research-related papers should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published only by invitation.

Manuscript. TEX and LATEX are preferable, standard Microsoft Word format (.doc) is acceptable. The author's JTIT LATEX style file is available:
http://www.itl.waw.pl/publ/jtit/doc/jtitaut.zip

Papers published should contain up to 10 printed pages in LATEX author's style (Word processor one printed page corresponds approximately to 6000 characters).

The manuscript should include an abstract about 150–200 words long and the relevant keywords. The abstract should contain statement of the problem, assumptions and methodology, results and conclusion or discussion on the importance of the results. Abstracts must not include mathematical expressions or bibliographic references.

Keywords should not repeat the title of the manuscript. About four keywords or phrases in alphabetical order should be used, separated by commas.

The original files accompanied with pdf file should be submitted by e-mail: redakcja@itl.waw.pl

Figures, tables and photographs. Original figures should be submitted. Drawings in Corel Draw and PostScript formats are preferred. Figure captions should be placed below the figures and can not be included as a part of the figure. Each figure should be submitted as a separated graphic file, in .cdr, .eps, .ps, .png or .tif format. Tables and figures should be numbered consecutively with Arabic numerals.

Each photograph with minimum 300 dpi resolution should be delivered in electronic formats (TIFF, JPG or PNG) as a separated file.

References. All references should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. Samples of correct formats for various types of references are presented below:

[1] Y. Namihira, "Relationship between nonlinear effective area and mode field diameter for dispersion shifted fibres", *Electron. Lett.*, vol. 30, no. 3, pp. 262–264, 1994.

[2] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

[3] S. Demri and E. Orłowska, "Informational representability: Abstract models versus concrete models", in *Fuzzy Sets, Logics and Knowledge-Based Reasoning*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.

Biographies and photographs of authors. A brief professional author's biography of up to 200 words and a photo of each author should be included with the manuscript.

Galley proofs. Authors should return proofs as a list of corrections as soon as possible. In other cases, the article will be proof-read against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within one week after receiving the offprint.

Copyright. Manuscript submitted to JTIT should not be published or simultaneously submitted for publication elsewhere. By submitting a manuscript, the author(s) agree to automatically transfer the copyright for their article to the publisher, if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and all translation rights. No part of the present JTIT should not be reproduced in any form nor transmitted or translated into a machine language without prior written consent of the publisher.
For copyright form see:
http://www.itl.waw.pl/publ/jtit/wskazowki.html

A copy of the JTIT is provided to each author of paper published.

Journal of Telecommunications and Information Technology has entered into an electronic licencing relationship with EBSCO Publishing, the world's most prolific aggregator of full text journals, magazines and other sources. The text of *Journal of Telecommunications and Information Technology* can be found on EBSCO Publishing's databases. For more information on EBSCO Publishing, please visit www.epnet.com.

INSTYTUT ŁĄCZNOŚCI
PAŃSTWOWY INSTYTUT BADAWCZY

Editorial Office

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

tel. +48 22 512 81 83
fax: +48 22 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.itl.waw.pl/jtit