


e – Society Journal

Research and Applications

Volume 3 Number 1
July, 2012



E-society Journal

Research and applications

Volume 3, Number 1, July 2012

ISSN 2217-3269

COBISS.SR-ID 255833863

E-society Journal

Research and applications

Publisher:

University of Novi Sad

Technical faculty "Mihajlo Pupin"

Djure Djakovica bb, Zrenjanin, Serbia

Editor:

Miodrag Ivković

University of Novi Sad

Technical faculty "Mihajlo Pupin" Zrenjanin

CIP – Каталогизација у публикацији
Библиотека Матице српске, Нови Сад

621.3

004.4

E-society Journal: research and applications / editor Miodrag Ivkovic.

- Vol 3. No. 1 (jul. 2012) - . – Zrenjanin: University of Novi Sad,

Technical faculty "Mihajlo Pupin", 2012-. – 25 cm

Dva puta godisnje.

ISSN 2217-3269

COBISS.SR-ID 255833863

Printed by: PC centar Magus, Zrenjanin

Printing: 50 copies

Zrenjanin, 2012.

Contents

The Current Design of E-speranto and the Evaluation of its Interpretation in Slovenian	1
Grega Jakus, Sašo Tomažič	
Agent Based Data Mining in Wireless Sensor Networks: A Survey	11
Milica Knežević, Nenad Mitić, Zoran Ognjanović, Veljko Milutinović	
CoAP (Constrained Application Protocol) implementation in M2M Environmental Monitoring System	21
Tomislav Dimčić, Srđan Krčo, Nenad Gligorić	
Empirical Findings of the Awareness and Application of Barcodes in Austria	35
Iris Uitz, Michael Harnisch, Bernd M. Zunk	
Modeling and Navigation of an Autonomous Quad-Rotor Helicopter	45
Gyula Mester, Aleksandar Rodic	
Clustering Multiple Datasets Under Parameter Similarity Constraints	55
Nikola S. Milosavljević, Dušan Đ. Okanović	
Realistic Terrain Aware Mobility Model	67
Maja Dineska, Sonja Filiposka	
Simulating a Two-Stage Packet Scheduler	79
Anton Kos, Sašo Tomažič	

The Current Design of E-speranto and the Evaluation of its Interpretation in Slovenian

Grega Jakus, Sašo Tomažič

Faculty of Electrical Engineering, University of Ljubljana, Slovenia

Abstract—The present paper describes the current design of E-speranto, a formal language for recording multilingual documents on the Web. The message in E-speranto can be represented with a semantic network which consists of concepts, the relations between concepts and concept attributes. The concrete syntax of E-speranto is based on XML (eXtensible Markup Language), enabling a simple addition of multilingual documents recorded in E-speranto into web pages. When a user visits such multilingual web page, the interpreter interprets the content in E-speranto into a preferred natural language. In this way, the user can read multilingual web documents in any chosen language, provided that the language is supported by the interpreter.

Besides the design of E-speranto, the paper also presents the results of a quantitative evaluation of E-speranto interpretation in Slovenian. The evaluators included in the evaluation were potential system users and professional translators. The concept of the interpretation of multilingual web documents was compared to the existing approach to multilingual Web, i.e. the use of online translation tools. The results of the comparison prove the adequacy of the concept of multilingual Web based on E-speranto, as the interpretation yielded significantly more accurate sentences than the online translation tool used in the evaluation.

Keywords—E-speranto, interpretation, World Wide Web, XML

I. INTRODUCTION

Multilingualism appears on the World Wide Web in two ways. Most often it is reflected in the translations of web pages into languages which, in the owners' opinion, cover most of their target audience. The users that do not understand any of these languages can use one of the freely available online translation tools, such as Babel Fish [1], Google Translate [2], PROMT [3] or SYSTRANet [4]. The most common weaknesses of these tools are a lower quality of translations and a limited scope of supported languages. This is mostly due to two of the most prominent problems in the field of machine translation, namely, natural language understanding and the problem of scalability.

The natural language understanding is demanding due to the ambiguities and inconsistencies that are present in natural languages. The problem of scalability is especially apparent when translating between a large number of languages. In order to translate between n languages, we require $n(n-1)$ translation tools, which actually means that 47,727,372 translators are needed in order to provide the translation between the 6909 world's known living languages [5].

The established approach to address the problem of scalability is a translation in two separate steps via an intermediate language or interlingua [6]. The interlingua approach reduces the necessary number of modules to $2n$ but still it does not resolve the issues with natural language understanding. As this is problematical only when translating from a natural language into an interlingua and not in the reverse process, the solution lies in the introduction of a

formal language for recording multilingual documents. Such a language would reduce the need for automatic translation from a natural language, as it would enable the authors to create documents using specially designed tools. E-speranto was designed especially for this purpose. When translators into E-speranto are developed, this language will also function as an interlingua in multilingual translation.

The idea and the early design of E-speranto can be found in [7] and [8]. In this paper, we present the current design of E-speranto which is the result of the progress in the development of this language.

II. RELATED WORK

E-speranto belongs to abstract formal languages intended for multilingual communication. Already established examples of such languages are the so-called interlinguae. The record in an interlingua is the result of a lexical, structural and semantic analysis of a text in a natural language. As the analysis removes all specificities of the source language, the record in an ideal interlingua is independent of any natural language and thus completely abstract, expressing the meaning only. Due to this complete abstraction, the interlingua can serve as the basis for the generation of content in any given natural language. Some more notable implementations are presented in the next paragraphs.

DLT (Distributed Language Translation) [9] was a project of the development of a multilingual translation system in the 1980s that used an adapted version of Esperanto as an interlingua. The document in Esperanto would be carried over the computer network and interpreted in a chosen language by the target computer. Although DLT presented a novel approach to machine translation, it was established that Esperanto is not suitable for an interlingua, as this language still has many features similar to those of natural languages.

The interlingua of the system KANT [10] is based upon controlled English (a language with a limited scope of vocabulary) and was created with the intention of translating technical documentation. Its interpretation produces very accurate sentences, but due to the limited field of use it is not directly applicable for general multilingual translation.

The interlingua of the Rosetta system [11] inherently contains the features of supported languages, which significantly simplifies its interpretation. However, as it contains only the linguistic features of a limited set of languages, the interlingua is not universal and the system itself is not scalable, as we would need to change the interlingua if we decided to add a new language into the system.

UNL (Universal Networking Language) [12] is a computer language for recording and exchanging information in the Internet. UNL is a successor of the ATLAS-II [13] and PIVOT [14] interlinguae. The authors are able to write in UNL directly, but due to the fact that the language itself was not intended for such use, it is not easily comprehensible.

III. E-SPERANTO

The development of E-speranto was influenced by Esperanto, a language designed at the turn of the 20th century. The aim of Esperanto was to become the international auxiliary language. The most important features of Esperanto are the relative consistency and unambiguousness of its grammar and a suitable level of expressiveness. Although it has never

become a universal second language, Esperanto encompasses an extensive linguistic knowledge and years of work, which is why it was used as the basis for the development of E-speranto. The latter can thus be considered as an electronic version of Esperanto.

A. The message in E-speranto

The basic elements of the message in E-speranto are concepts, the relations between concepts and concept attributes. The message in E-speranto can be represented with a semantic network (Fig. 1).

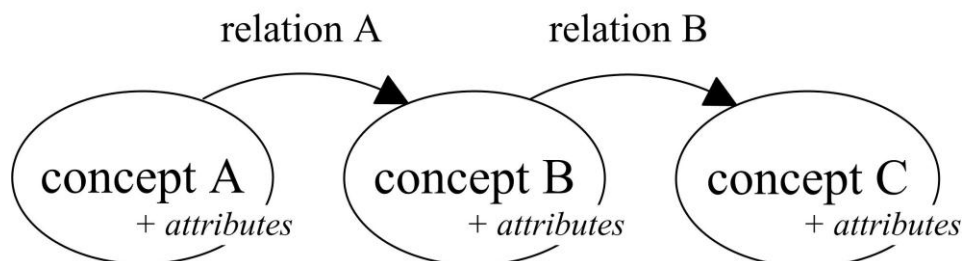


Fig. 1 The model of the message in E-speranto

A concept is an abstract idea or a symbol and can also be defined as a unit of meaning. One of the important challenges when designing a language intended for multilingual communication is the choice of its vocabulary, as it defines the accuracy of expression and the expressiveness of the language. When developing the basic vocabulary of E-speranto, we used the vocabulary of Esperanto together with the meanings that are conveyed by the individual items in the vocabulary.

In addition to the concepts, the relations between the concepts also have an important contribution to the meaning of the message. The relations in E-speranto can be divided into logical and semantic. The basis for determining the set of relations is the grammar of Esperanto. However, as it turned out that the relations from Esperanto do not suffice for a suitable semantic description of the content, the set of relations is continuously supplemented.

The attributes of the concepts transform the concepts into concrete objects by placing them in the world as it is perceived by the author of the message. The attributes can for example refer to the time (past, present, future, not stated etc.), aspect (finite, non-finite), mood (conditional, imperative, indicative) etc. Whereas some attributes are vital for the retention of meaning when interpreting E-speranto in a natural language, the other attributes only indicate the style with which the author has formed his message. As is the case with semantic relations, the grammar of Esperanto is the basis for defining the set of attributes and their values; however, new attributes and their values can be added if required.

B. Concrete syntax

The concrete syntax of E-speranto is based on the syntax of XML (eXtensible Markup Language). The latter is compatible with XHTML (eXtensible HyperText Markup Language), which enables the integration of documents written in E-speranto into web pages. The grammar rules of E-speranto are defined using the XML Schema language [15]. The basic standalone

unit in the concrete syntax is the element *sentence* which roughly corresponds to a sentence in a natural language.

C. Concepts

The concepts that constitute a sentence in E-speranto are recorded using lexical units from the vocabulary of Esperanto and are then placed in the element *concept*. Each concept is assigned to a class using a parent element of the element *concept* (Fig. 2).

The classes of concepts enable the distinction between the concepts as regards their role in the meaning of the message (Table I). Each class has its own grammatical rules defining the syntactic and semantic restrictions of the class. The restrictions, for example, determine the available subordinate classes, the set of attributes used to describe the concepts in more detail and the set of relations used to link a concept with other concepts.

D. Relations

A semantic relation is recorded as the value of the XML attribute *relation*. The direction of a relation is implied with the nesting of elements. The relation within a specific element implies the semantic relation of the concept in this element to the concept in a parent element (Table II).

```
<sentence feelings="declarative" organization="simple">
  <subject proper_name number="singular">
    <concept>e-speranto</concept>
  </subject>
  <predicate mood="indicative" tense="present" person="third">
    <concept>esti</concept>
    <predicate detail_predicate="predicate_noun" number="singular">
      <concept>dezajno</concept>
      <object relation="composition_element" number="singular">
        <concept>lingvo</concept>
        <attribute detail_attribute="relativity">
          <concept>formala</concept>
        </attribute>
      </object>
    </predicate>
  </predicate>
</sentence>
```

Fig. 2 The record of a sentence in E-speranto

(DEZAJNO = “design”, LINGVO = “language”, ESTI = “to be (an instance of something)”, FORMALA = “formal”)

TABLE I
THE CLASSES OF CONCEPTS

Class	Interpretation
predicate	The concept that represents an action or state.
subject	The concept with the semantic role of the doer of the action.
object	The concept that is involved in the action or state.
adverbial	The concept describes the circumstances of the action or state.
attribute	The concept describes the features of the other concepts.

TABLE II
THE RECORD OF SEMANTIC RELATIONS

Semantic relation	Record in E-speranto
“A is the recipient of B”	<pre><predicate> B <object relation="recipient"> A </object> </predicate></pre>
“A is an instrument for B”	<pre><object> B <object relation="instrument"> A </object> </object></pre>
“A is a part of B”	<pre><object> B <object relation="partOf"> A </object> </object></pre>

TABLE III
THE RECORD OF LOGICAL RELATIONS. (THE ELEMENT X REPRESENTS ANY OF THE AVAILABLE CONCEPT CLASSES.)

Semantic relation	Record in E-speranto
“A and B”	<pre><sequence relation="and"> <x> A </x> <x> B </x> </sequence></pre>
“A or B”	<pre><sequence relation="or"> <x> A </x> <x> B </x> </sequence></pre>

The available set of semantic relations depends on which class the concept is assigned to. The concepts in the class *subject* have a predetermined role of the doer of the action represented by a concept belonging to the class *predicate*. The class *object* allows only the use of semantic relations linking nominal concepts. The use of other relations (temporal, locative, causal etc.) is only allowed with the concepts assigned to the class *adverbial*.

Semantic relations can be combined with logical relations by using the element *sequence* (Table III). As opposed to semantic relations, logical relations do not depend on the class of the concept.

E. Attributes

The concept attributes are recorded in the form of XML attributes (Table IV). As is the case with semantic relations, concept attributes also depend on the class to which the concept has been assigned. For example, tense or aspect can be assigned only to the concept belonging to the class *predicate*.

TABLE IV
THE EXAMPLES OF CONCEPT ATTRIBUTES

Semantic relation	Record in E-speranto
aspect of the activity	<predicate aspect="ongoing">
plural	<subject number="plural">
biological gender	<object gender="female">

IV. THE PROOF-OF-CONCEPT SYSTEM

The system for the interpretation of E-speranto consists of prototype interpreters for Slovenian and English and the related information resources which mostly contain linguistic data. Due to the demanding and extensive development of the system, we decided on several simplifications which are mostly reflected in the limited functionality of the interpreters and content of the information resources, as well as in the simplified record in E-speranto.

The prototype E-speranto interpreters and the information resources were used to design a proof-of-concept implementation of the multilingual Web based on E-speranto (Fig. 3). The central elements of the proof-of-concept implementation are the development environment for the record of documents in E-speranto and the multilingual web site. The latter contains multilingual web documents and also the system for their interpretation. As the server responds the request for a multilingual document with a usual XHTML document with the content in a chosen target language, the user can test the multilingual Web based on E-speranto by using any web browser. The system can be tested on the E-speranto web site [16].

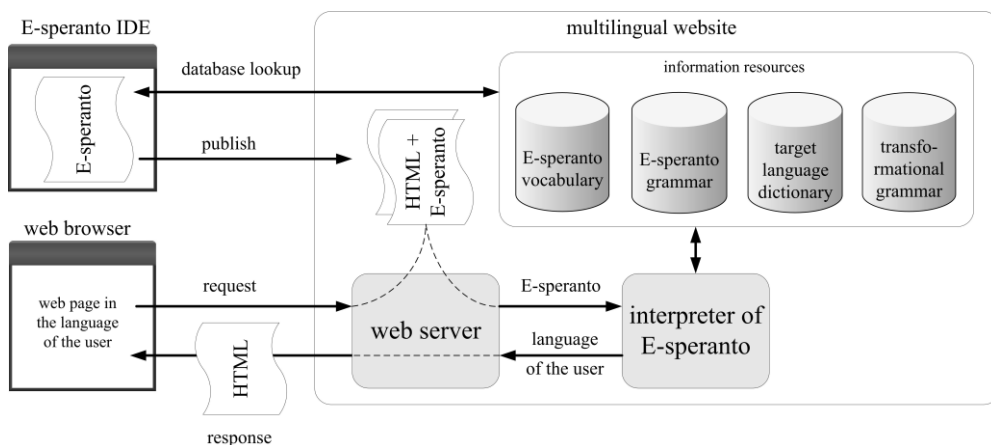


Fig. 3 The proof-of-concept implementation of the multilingual Web based on E-speranto

V. EVALUATION

We concluded a quantitative evaluation of the E-speranto interpretation by using the proof-of-concept system. For this purpose, we chose four English texts which altogether consist of 46 simple sentences. The selected texts include a welcome message from a mobile operator to the roaming user and three recipes.

Due to the limited functionality of the interpreters, the texts were adapted before recorded in E-speranto. The information resources were also supplemented by the records required for the interpretation, however, the functionality of the interpreters remained unchanged. The record in E-speranto was then interpreted into Slovenian using the prototype interpreter, while the adapted original text was at the same time translated with the online translation service Google Translate.

The evaluation was undertaken in two parts. In the first part, ten potential users of the system, all Slovenian native speakers, were given both the translated and the interpreted texts. The evaluators were not shown the original text and were also not revealed the manner in which each text was created. The evaluators were asked to mark each sentence in the translated and the interpreted texts with marks from 1 ("very bad") to 5 ("perfect"). The subject of the evaluation included the clarity of the expressed meaning and grammatical correctness.

The second part of the evaluation was conducted in a similar way, the only difference being that, instead of potential users, the evaluation was conducted by three qualified translators. They were asked to compare the translated and the interpreted texts with the adapted original text and evaluate if they contain the same meaning.

The average marks of the translation using the online translation service Google Translate and the interpretation using the prototype interpreter of E-speranto are given in Table V.

TABLE V
THE RESULTS OF THE EVALUATION

	precision of expression		grammatical correctness		preservation of meaning	
	Google Translate	interpreter of E-speranto	Google Translate	interpreter of E-speranto	Google Translate	interpreter of E-speranto
welcome message	4,82	4,73	4,95	4,27	4,72	4,72
recipe 1	3,87	4,64	3,42	4,59	3,86	4,45
recipe 2	3,96	4,93	3,41	4,89	3,92	4,56
recipe 3	3,60	4,91	3,18	4,76	3,43	4,83
all texts	3,93	4,81	3,54	4,68	3,86	4,63

VI. DISCUSSION

According to the results of the evaluation, the interpretation of E-speranto proves to be more accurate than the translation with the online translation service Google Translate in all three observed aspects. However, if we were to repeat the evaluation by choosing a random text, the results would most probably turn in favor of the Google Translate. As the information resources in the system for interpreting E-speranto are limited, it is unlikely that they would contain the exact content required for the interpretation of the selected text. The results given in Table V should thus not be interpreted as a comparison between the prototype interpreter of E-speranto and an existing translation tool, but rather as a comparison between two approaches to multilingualism on the Web.

The results of the evaluation reveal the potential of multilingual Web based on E-speranto. The key strength of the approach is a very precise interpretation of E-speranto, as very few words turn out to be wrongly translated or not translated at all. This is due to the fact that information resources in the system for the interpretation can easily be supplemented. Although the vocabulary and set of semantic relations are currently limited, they are being regularly supplemented with further development of E-speranto and its interpreters.

As a part of the burden for a precise interpretation falls on the user, the establishment of E-speranto mostly depends on the complexity of the recording in this language. This is why special attention has been dedicated to creating a development environment for recording texts in E-speranto. Although the existing environment proved helpful when recording in E-speranto, it will need to be upgraded. The development of graphical editors would, for example, make the recording in E-speranto more intuitive and the users would no longer be required to learn the exact syntax of E-speranto.

VII. CONCLUSION

In this paper, we briefly presented the current design of E-speranto and the evaluation of its interpretation in Slovenian. The interpretation of E-speranto was compared to the use of an online translation tool. The results of the comparison prove the adequacy of the concept of multilingual Web based on E-speranto, as the interpretation yielded more accurate sentences

than the translation tool used in the evaluation.

As a part of the future development, we intend to supplement the vocabulary of E-speranto and the set of semantic relations. In addition, we intend to evaluate the interpretation of E-speranto based on a larger number of texts and develop the interpreters also for other Slavic languages.

REFERENCES

- [1] Yahoo! Babel Fish - Text Translation and Web Page Translation, <http://babelfish.yahoo.com>, accessed 12. January 2012.
- [2] Google Translate, <http://translate.google.com>, accessed 12. January 2012.
- [3] PROMT Translation Software and Dictionaries, <http://www.promt.com>, accessed 12. January 2012.
- [4] SYSTRANet - Online translation software and tools, <http://www.systranet.com>, accessed 12. January 2012.
- [5] Paul, L. M. (ed.), *Ethnologue: Languages of the World*, Sixteenth edition, SIL International, Dallas, 2009.
- [6] Hutchins, W. and Somers, H., *An Introduction to Machine Translation*, Academic Press, New York, 1992.
- [7] Tomažič, S., Multilingual Web with E-speranto, *IPSI BgD Transactions on Internet Research*, Vol. 3, No. 2, pp. 13-15, 2007.
- [8] Omerović, S., Jakus, G., Filimonova, T. and Tomažič, S., Zapis večjezičnih besedil v e-sperantu, *Electrotechnical Review*, 74 (4), pp. 151-157, 2007.
- [9] Schubert, K., The Architecture of DLT – interlingual or double-dialect, in: Maxwell, D., Schubert, K. and Witkam, T. (eds.), *New Directions in Machine Translation*, Floris Publications, Holland, 1988.
- [10] Nyberg, E. and Mitamura, T., The KANT system: Fast, accurate, high-quality translation in practical domains, in *Proceedings of the 14th conference on Computational linguistics*, 1992.
- [11] Rosetta, M. T. (pseud.), *Compositional Translation*, Kluwer Academic Publishers, Dordrecht, 1994.
- [12] Uchida, H., Zhu, M. and Della Senta, T., *Universal Networking Language: A gift for a millennium*, The United Nations University, Tokyo, 1999.
- [13] Uchida, H., *ATLAS II: A Machine Translation System Using Conceptual Structure as an Interlingua*, in *Proceedings of the Second Machine Translation Summit*, Tokyo, 1989.
- [14] Muraki, K., *PIVOT: Two-Phase Machine Translation System*, in *Machine Translation Summit: Summit Manuscripts and Program*, Japan, 1987.
- [15] Thompson, H. S., Beech, D., Maloney, M. and Mendelsohn, N. (eds.), *XML Schema Part 1: Structures*, W3C Recommendation 28 October 2004, <http://www.w3.org/TR/xmlschema-1/>, accessed 12. January 2012.
- [16] E-speranto web page, <http://www.e-speranto.org>, accessed 12. January 2012.

Agent Based Data Mining in Wireless Sensor Networks: A Survey

Milica Knežević*, Nenad Mitić**, Zoran Ognjanović*, Veljko Milutinović***

* Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia

** School of Mathematics, University of Belgrade, Serbia

*** School of Electrical Engineering, University Belgrade, Serbia

Abstract—Wireless Sensor Networks (WSNs) are nowadays widely used for monitoring physical phenomena of the environment. Traditional data mining algorithms cannot be applied to such distributed systems in which strict memory, power, and communication constraints do exist. New algorithms that would be more suitable for WSNs have been developed. In this paper, we give an overview of the selected algorithms for data mining in WSNs and discuss benefits of integration of agent systems and data mining algorithms.

Keywords—Wireless sensor networks, Data mining, Software agents.

I. INTRODUCTION

The WSNs consist of a large number (typically hundreds or thousands) of lightweight devices - sensor nodes that collect measurements from the environment. Sensor nodes are battery powered and equipped with a wireless radio transceiver, small CPU, actuators, memory, and one or more sensors that can measure temperature, light, humidity, pressure, sound, vibration, etc. In general, they can be of one of the following types: physical, chemical, biological, or technological. There is a wide variety of use of WSNs in different fields like military surveillance, target tracking, traffic management, weather forecasting, habitat monitoring, smart home design, health and medical monitoring, etc.

A. Data Mining in Wireless Sensor Networks

Data Mining (DM) is a discipline of examining large amounts of data in order to find unseen patterns and relations. In WSNs, data mining techniques can be used for noise, event and attack detection, target classification, energy conservation, etc. Traditional data mining approaches assume that all data are static and stored at some central repository (often organized as a database). Collecting all sensors data at some central point leads to the bottleneck problem, increase of latency, and greater energy overhead. Performing some part of the computation at sensor nodes and sending the processed and aggregated data instead of raw data is more energy efficient and can extend the lifetime of the network. This leads to development of new algorithms that can extract data patterns from sensors data in a more efficient way.

B. Agent Systems

Agent is a software unit capable of independent action on behalf of its user or owner. Once created, agent does not require interaction with user and can activate itself not strictly only being invoked for a task. Agent may reside on a host waiting for the starting conditions. Agents may also communicate and collaborate with each other while performing tasks. Characteristics

of an agent, which distinguish it from an arbitrary program, are: (a) persistence, (b) autonomy, (c) reactivity, and (d) social ability.

Mobile agents are a special kind of agents that can migrate, under host or its own control, from one node of a network to another. A mobile agent running at a node can suspend its execution at some point, move to another host, and resume execution from the suspension point.

Agents can be found useful for data mining over WSNs for the following reasons:

- Agents react to the changes in the environment and are able to bypass the faulty host or select a more reliable host.
- Agents can cooperate, negotiate, collaborate, and form a smart system.
- Mobile agents allow algorithm to scale well. The parent agent can clone several child agents to implement concurrent operations.
- Mobile agents operate asynchronously. Once dispatched, mobile agent executes autonomously without any intervention of the dispatcher.

II. PROBLEM STATEMENT

The purpose of this paper is to: (a) identify important criteria for classification of data mining techniques for WSNs, (b) describe classification criteria and present a classification tree, (c) give a description of each one of the selected algorithms, and (d) show how agent systems can help improving DM algorithms to suit distributed nature of WSNs, as well as communication and computational constraints.

III. CLASSIFICATION CRITERIA AND CLASSIFICATION TREE

Possible classification criteria for DM methods in WSNs are presented as a tree (see Fig. 1). The first level of the tree divides methods into two groups: Collaborative and Local-to-Global Learning approach.

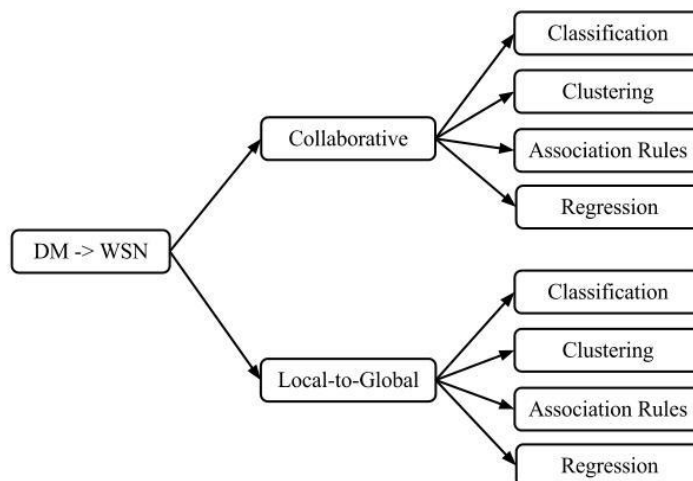


Fig. 1 The proposed classification tree

In the Collaborative Learning approach (see Fig. 2), each sensor node creates a data mining model exchanging multiple rounds of messages with its neighboring nodes. Messages may consist of a selected subset of aggregated local data or, more often, parameters of the local models. This approach does not require hierarchical organization of the network. Stationary agents may be placed at nodes, autonomously selecting other agents to collaborate in building the data mining model.

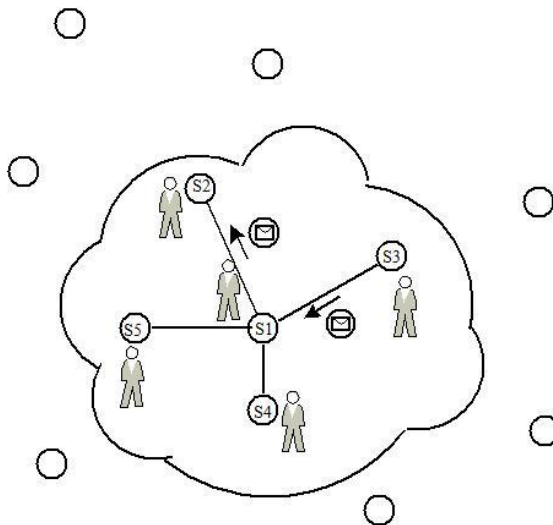


Fig. 2 Collaborative Learning – Stationary agents placed at sensor nodes exchange multiple rounds of messages with neighboring agents in order to create data mining model.

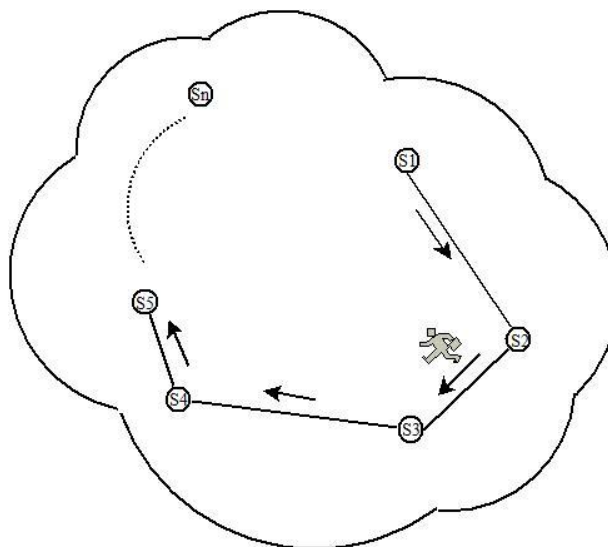


Fig. 3 Local-to-Global Learning – A mobile agent sent to visit sensor nodes, collects local models. When the last node S_n is visited, global model can be created.

In the Local-to-Global Learning approach (see Fig. 3), each sensor node builds its local model. The global model is created as a combination of local models. This approach usually requires a hierarchical organization of the network. Data mining models created at child nodes are sent to be combined with parent data model, until the root node is reached. In order to avoid this requirement, a mobile agent, or several of them, may be dispatched into the network, to collect and combine local models into a global model.

On the second level of the tree, the appropriate DM algorithms for each one of the mentioned groups are presented. More details about used algorithms are provided in the next sections of this paper.

IV. PRESENTATION OF THE EXISTING ALGORITHMS

A. Collaborative Learning

1) Classification

An algorithm for unsupervised classification based on the one-class quarter-sphere Support Vector Machine (SVM) is proposed in [1]. This is an algorithm for outlier detection in real time. Each node builds its local model and exchanges the model parameters with its spatially neighboring nodes. Sensor node will use collected parameters to distinguish between event and error, which are common outliers, in the process of classification, yielding a lower false alarm rate. Nodes in the WSN are considered to be time synchronized and densely deployed and sensors data correlated in time and space.

Learning a one-class quarter-sphere SVM model is a linear optimization problem and its computational complexity is reduced compared to some other one-class SVM-based classifiers. The algorithm scales well to large networks. The solution for model update is currently under development. Further work will include improvements towards robustness in presence of network topology changes.

Algorithm One-Class Quarter Sphere SVM

procedure LearnLocalSVM:

for each node N_i :

 collect m data measurements;

 learn local quarter-sphere radius R_i ;

 send local radius to the neighboring nodes;

 collect radii from the neighboring nodes;

 compute R'_{im} median radius of the collected radii;

 compute R_{im} median radius of the collected radii and R_i ;

 call **IsOutlier**;

procedure IsOutlier:

when new measurement x_i arrives

 compute d_i as distance of x_i from the quarter-sphere center

 if ($d_i > R_i$ AND $d_i > R'_{im}$)

x_i indicates an outlier;

 call **SourceOfOutlier**;

 else x_i indicates normal measurement;

procedure SourceOfOutlier:

let $x_{i1}, x_{i2}, \dots, x_{ik}$ be new data measurements arriving at N_i 's neighboring nodes;
 collect d_{i1}, \dots, d_{ik} distances of the new data measurements from their quarter-sphere centers;
 compute d'_{im} as median distance of collected distances
 if ($d_i > R_i$ AND $d'_{im} > R'_{im}$)
 if ($d_i > R_{im}$ AND $d'_{im} > R'_{im}$)
 x_i may indicate an event;
 else x_i may indicate an erroneous measurement;

2) Clustering

A Peer-to-Peer version of the K-Means clustering algorithm is proposed in [2]. The algorithm does not require global synchronization, i.e. not all the nodes have to be in the same iteration step, but there is an upper bound and the algorithm is not completely asynchronous. In order to create accurate data models, sensor nodes communicate with their immediate neighbors exchanging locally produced cluster summaries (centroids and the number of data vectors per cluster).

Links inside the network may come up and down but the nodes will continue building the model exchanging messages with their available neighbors. The authors propose an idea for handling non-static data, but without experimental verification.

Algorithm P2P K-Means

for each node N_i :
 initialize K centroids;
 repeat // once t time units have elapsed
 assign local points to the K centroids;
 update local centroids;
 if centroids changed significantly
 set changed = true;
 else set changed = false;
 poll the collection of nodes for their centroids, cluster counts and changed flag;
 process polling requests received from other nodes;
 combine local centroids with the received ones;
 until change flag is false for N_i and nodes that responded;

3) Association Rules

An algorithm for mining association rules between spatial-temporal events detected by sensor nodes is proposed in [3]. Correlated events often occur in spatial and/or temporal proximity, thus each node periodically collects notifications about the events that are detected by the nodes within a certain distance. To reduce memory consumption during frequent item set extraction, two approaches are considered. The first approach is the approximate mining of frequent itemsets. The quality of approximation will depend on the available memory. The second approach is to find only a subset of frequent itemsets (close or maximal frequent itemsets).

Paper [3] shows that in-network creation of association rules in systems like WSNs, where strict constraints on memory and computational power do exist, is a challenging problem and that more energy efficient algorithms, which can be applied in practice, should be developed.

Algorithm Distributed Station-Temporal Event Patterns

for each node N_i :

- collect information about local events;
- periodically collect information about events from the neighboring nodes;
- extract frequent/closed/maximal itemsets;
- create assoc. rules using extracted itemsets;

B. Local-to-Global Learning

1) Classification

A distributed version of C4.5 classification algorithm is proposed in [4]. The paper also proposes the routing protocol that will provide hierarchical organization of sensor nodes into a spanning tree. Local classifiers are built starting from the leaf sensor nodes and merged along the routing path. In order to create the enhanced classifier from the downstream classifiers, the downstream node will restore pseudo training data set for each received classifier. The role of pseudo data set is to reflect the distribution of different classes.

In order to avoid immediate propagation of the model update from the leaf node, over intermediate nodes, up to the root node, a parent node may decide to periodically check for changes of the model at the child nodes. This approach is more energy efficient. The algorithm gives high classification accuracy with low storage and communication overhead. The algorithm does not take into consideration node failure and packet loss.

Algorithm Enhanced C4.5

for each node N_i :

- if node is a leaf
 - periodically collect data;
 - create local classifier;
 - send local classifier to the parent;
- else
 - periodically collect data;
 - for each new classifier from child
 - generate pseudo data set;
 - create combined classifier from local data and pseudo data;
 - if not root node send local classifier to the parent;

Two algorithms for incremental training of a SVM based classifier are proposed in [5]. The second algorithm is a modification of the basic algorithm and is suitable for networks where regular updating of the model is necessary (problem known as *concept drift*). The authors accepted the protocol for spatial clustering of WSN, which reduces the amount of data sent through the network. Starting from the first cluster head, local estimation SVM is built and sent to the next cluster head where the enhanced estimation SVM is created using local sample vectors and the previous estimation.

The experiments show that both algorithms give a good approximation of the model created using the centralized approach while reducing energy costs by more than 50%.

Algorithm Incremental SVM Training

```

for each cluster head  $ch_i$  in  $\{ch_1, \dots, ch_n\}$ :
    create local estimation  $SVM_i$  from local data at  $ch_i$  and previous estimation  $SVM_{i-1}$ ;
    if  $ch_i \neq ch_n$ 
        send model to the next cluster head;

```

2) Clustering

An algorithm for anomaly detection based on fixed radius clustering is presented in [6]. The algorithm requires nodes to be time synchronized and to maintain tree network topology in order to produce global clustering model at the root node. Clusters which include outliers are detected using average inter-cluster distance of the K nearest neighbors.

Statistics of the local clusters consists of the linear sum of data vectors and the number of data vectors in the cluster. The experimental results show that the cluster radius has to be selected carefully in order to achieve better detection performance. A proper value can be selected by training the system before deployment.

Algorithm Fixed Radius Distributed Clustering

```

for each node  $N_i$ :
    if leaf node
        create  $K$  local clusters;
        send statistics for local clusters to the parent;
    else if not root node
        create  $K$  local clusters;
        merge local clusters and clusters received from children;
        send statistics for clusters to the parent;
    else
        create  $K$  local clusters;
        merge local clusters and clusters received from children;
        identify clusters with outliers;

```

An algorithm for kernel density based clustering, with two possible agent-oriented implementations, is proposed in [7]. All nodes create local estimates using same kernel function, a window width parameter that controls smoothness of the estimate, and same sampling parameters. Density estimates are additive for homogeneously distributed data and local estimates can be collected and summed up giving the global estimate. Multi-dimensional sampling is used to avoid explicit reference to data objects.

Possibility of efficient clustering over all data in a WSN is not discussed in paper [7]. The global density estimate is created over all data, but it is only used to create local clusters at each node.

Algorithm Kernel Density Estimate Clustering

```

for each node  $N_i$ :
  if  $N_i$  is not helper node
    create samples of the local density estimate;
    send local samples to the helper node;
    when samples of global estimate are received
      find  $M=\{M_1...M_n\}$  local maximums of the global estimate;
      for each data vector  $v_i$ :
        if  $v_i$  can be connected using an uphill path with  $M_j$ 
          assign  $v_i$  to the cluster with center  $M_j$ ;
  else
    collect local density estimates;
    create global estimate summing up local estimates;
    send global estimates to the nodes;

```

3) Regression

An algorithm for distributed kernel regression is proposed in [8]. At the first sensor node, kernel functions and coefficients of a regression model are selected based on local data and sent to the next sensor node. If the next visited sensor node is in the neighborhood of any previously visited sensor nodes, its kernel function may be well approximated by a linear combination of the model kernel functions. Otherwise, model order should be incremented by including node's kernel function. In both cases, coefficients of the model will be updated.

The algorithm can be applied in densely deployed sensor networks and it will still show low computational complexity, unlike many other regression algorithms.

Algorithm Reduced Order Kernel Regression

```

for each node  $N_i$  in  $\{N_1, \dots, N_m\}$ :
  if  $N_i$  is not in the neighborhood of previously visited nodes
    include corresponding kernel function to the model;
  update the coefficient vector to be as close as possible to the previous one;
  if  $N_i \neq N_m$ 
    send model to the next node;

```

V. CONCLUSIONS

In this paper we have proposed a classification of the existing distributed data mining algorithms for WSNs and possible integration of agent systems with each class of algorithms. This paper is intended for those who work with WSNs and are interested in energy efficient in-network data mining for WSNs. The idea that lies behind presented algorithms is that computation can be done in-network by putting (some) computational load on sensor nodes themselves.

Agent based model does not always lead towards more efficient algorithm. Agents have to be lightweight or they will produce too much energy overload. The same undesirable result will be created when dispatching too many agents into the network. Efficient data mining algorithms that fully exploit the possibilities agent systems offer, are yet to be developed.

ACKNOWLEDGMENT

This work was supported by the Serbian Ministry of Education and Science (project III44006).

REFERENCES

- [1] Y. Zhang, N. Meratnia, and P. Havinga, "An Online Outlier Detection Technique for Wireless Sensor Networks using Unsupervised Quarter-Sphere Support Vector Machine," Proceedings of the 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2008), Sydney, Australia, 15-18 December 2008, pp. 151-156.
- [2] S. Bandyopadhyay, C. Giannella, U. Maulik, H. Kargupta, K. Liu, and S. Datta, "Clustering distributed data streams in peer-to-peer environments," Information Sciences, Volume 176, Issue 14, 22 July 2006, pp. 1952-1985.
- [3] K. Römer, "Distributed Mining of Spatio-temporal Event Patterns in Sensor Networks," <http://www.inf.ethz.ch/vs/publ/papers/roemer-eawms06.pdf>
- [4] X. Cheng, J. Xu, J. Pei, and J. Liu, "Hierarchical distributed data classification in wireless sensor networks," Computer Communications, Volume 33 Issue 12, July 2010, pp. 1404-1413.
- [5] K. Flouri, B. Beferull-Lozano, and P. Tsakalides, "Training a SVM-based Classifier in Distributed Sensor Networks," 14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006.
- [6] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. Bezdek, "Distributed Anomaly Detection in Wireless Sensor Networks," Proceedings of the IEEE Singapore International Conference on Communication Systems, 2006.
- [7] M. Klusch, S. Lodi, and G. Moro, "The Role of Agents in Distributed Data Mining: Issues and Benefits," IAT '03 Proceedings of the IEEE/WIC International Conference on Intelligent Agent Technology, 2003.
- [8] P. Honeine, M. Essoloh, C. Richard, and H. Snoussi, "Distributed regression in sensor networks with a reduced-order kernel model," Communications Society, 2008, pp. 1-5

CoAP (Constrained Application Protocol) implementation in M2M Environmental Monitoring System

Tomislav Dimčić*, Srđan Krčo*, Nenad Gligorić*

* Faculty of Organisational Science, Belgrade, Serbia

Abstract – In order to enable M2M devices to communicate over the existing Internet, Constrained RESTful Environments (CoRE) working group at the Internet Engineering Task Force (IETF) proposed Constrained Application Protocol (CoAP) as application layer protocol for IP based solutions. This paper presents a survey of the CoAP (Constrained Application Protocol), along with its implementation and evaluation in a real M2M system. The CoAP protocol was deployed in a mobile Environmental Monitoring System for transmission of a resource description and sensor environmental data from IoT nodes and vehicle tracking devices. The evaluation of the CoAP protocol was done by comparing the existing HTTP implementation with CoAP protocol in the same M2M communication scenario. Results shows that the performance of the CoAP compared to the HTTP based resource retrievals performs better in constrained environments and give possible direction for further development utilizing the latest advancements and the vision of the Internet of Things.

Keywords - M2M; CoAP; Wireless Sensor Networks; Internet Of Things, Implementation

I. INTRODUCTION

Over the past few years, Internet of Things (IoT) has emerged as one of the most interesting technologies of a Future Internet. Enabling a range of new applications, it is based on constrained networks, i.e. networks of devices with limited computing, memory and energy capabilities such as wireless sensor networks. A typical constrained devices are characterized by severe limits on throughput, available power, and particularly on the complexity that can be supported with limited code size and limited RAM per node. This has lead to a number of activities [1] aiming at defining communication protocols for such networks, while maintaining global connectivity.

In this paper we focus on the application level protocols. The existing protocols on this level, like HTTP, although considered simple from a regular computer or even smartphone perspective, are actually very complex and difficult to handle from the perspective of constrained devices [2].

CoRE working group has developed CoAP [3], a specialized web transfer protocol for use within constrained networks and nodes for machine-to-machine communication. This light-weight application layer protocol uses UDP protocol on transport layer and enables reliability mechanism on the highest layer. In addition, the CoAP protocol provides request/reponse interaction model that supports resource discovery, multicast, asynchronous message exchanges with low overhead and parsing complexity, simple proxy and caching capabilities. The CoRE working group adopted Representational State Transfer (REST) architecture approach. This architecture style together with the CoAP protocol empowers constrained networks with web service oriented architecture, providing similar functionality as the regular web services, but optimized for execution.

This paper discusses the CoAP protocol and its use for machine to machine (M2M) communication in vehicle tracking and environmental control application – in one type of constrained networks. The M2M devices used in this implementation are IoT nodes capable to monitor different environmental parameters such as NO₂, SO₂, CO₂, etc. GPRS is used to transfer data to the back-end servers located in the cloud [4]. At the moment, HTTP is used for transmission of the gathered data. In this paper we compare performance of such transfer in terms of the amount of data transferred and the time required to transfer data against CoAP utilization is the same deployment.

In section 2, CoAP protocol is discussed in detail. More about Telit GM682-GPS modems is presented in section 3. Section 4 describes our implementation of the CoAP protocol on the used devices. A comparison of the HTTP and the CoAP performance in the system implementation is given in section 5. Section 6 concludes the paper.

II. CONSTRAINED APPLICATION PROTOCOL

The use of web services in constrained networks is an important part of the emerging M2M communication paradigm. Therefore, the CoRE working group proposed CoAP to optimize the use of the RESTful web service architecture in constrained nodes and networks.

Utilization of available protocols like HTTP, FTP, SOAP, etc, is not possible in constrained networks due to limitations of the devices that comprise such networks. These protocols have large overheads and are using, from the constrained devices point of view, a heavy weight protocol on the transport layer – TCP (a 3-way handshake for connection and a 2-way handshake for disconnection). The large overhead reduces the size of the user data that can be transmitted (payload), as the maximum packet size for IEEE 802.15.4 is only 127 bytes. However, it is crucial that the application layer protocol enables integration with IP based networks, so an adequate solution had to be designed [5].

CoAP represents an application layer protocol that is compliant with the existing web protocols and and at the same time lowers the complexity for the constrained environment in which it is used.

CoAP is not just a compressed HTTP protocol. It takes a subset of the HTTP and optimises it for use in constrained networks. There is a number of features that enables devices with constrained resources to adjust to the existing web, such as built-in discovery, asynchronous message exchanges, UDP binding with optional reliability supporting unicast and multicast requests, low header overhead and parsing complexity, URI and Content-type support, simple proxy and caching capabilities, a stateless HTTP mapping, security binding to Datagram Transport Layer Security (DTLS).

A. CoAP model

CoAP utilize datagram-oriented transport such as UDP to asynchronously interchanges messages. Figure 1. represents a schematic view of layers used in this communication.

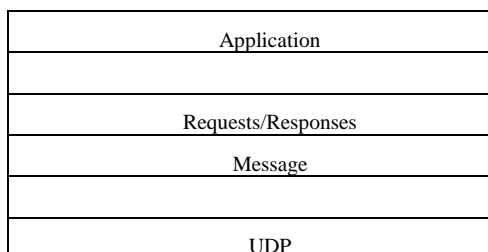


Figure 1. Abstract layering of CoAP

Although it looks like CoAP is logically using a two-layer approach, the Requests/Responses layer and the Message layer are just features of the CoAP header, and therefore it is a single layer protocol. The UDP layer receives a message from the message layer, packs it into a datagram and sends it to the IP layer of the OSI or the TCP/IP architecture. Similarly, in the opposite direction, the UDP layer receives a datagram from the IP layer and unpacks it into a message readable by the application layer.

Message Format

CoAP messages are encoded in a simple binary format. A message consists of a CoAP Header, options and a payload. Options are given in a Type Length Value (TLV) format. The number of options is determined by the header (Figure 2).

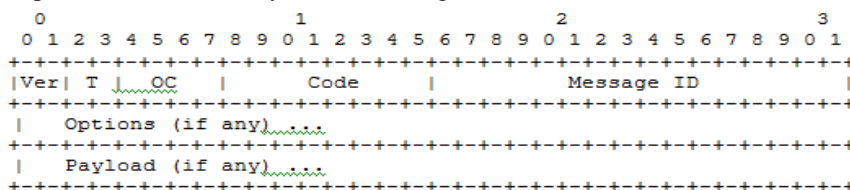


Figure 2. Message Format

The fields in the message are defined as follows:

Version (Ver): 2-bit unsigned integer. Indicates the CoAP version number. Currently, version is 1.

Type (T): 2-bit unsigned integer. Indicates if this message is of type Confirmable (0), Non-Confirmable (1), Acknowledgement (2) or Reset (3).

Option Count (OC): 4-bit unsigned integer. Indicates the number of options after the header.

Code: 8-bit unsigned integer. Indicates if the message carries a request (1-31) or a response (64-191), or is empty (0). (All other code values are reserved.)

Message ID: 16-bit unsigned integer. Used for the detection of message duplication, and to match messages of type Acknowledgement/Reset and messages of type Confirmable.

Options: Used for defining message type of the payload, Proxy-Uri, Uri-Host, Uri-Port, Location Path, Max-Age, Etag, Uri Path, Uri Query, Token, Accept, If-Match, etc.

Payload: The payload can carry the representation of a resource or some useful data from the sensor. Its format is specified by the Internet media type given by the Content-Type Option. Both requests and responses may include payload.

Message Type

Message type is determined by the code field in the header. Message can carry request, response or it can be empty.

Confirmable (CON) - message of this type requires response, i.e. confirm that the message arrived.

Non-Confirmable (NON) - this message does not require response, therefore it represents an unreliable message.

Acknowledgement (ACK) - message of this type is used to confirm that CON message is arrived. Message ID has to be the same as the CON ID.

Reset (RST) - message of this type is used to confirm that CON message is arrived but that there is something missing and the message could not be processed.

Reliability

For sleeping, low power devices, which has reduction of power consumption as one of the main design goals, reliability mechanisms that TCP offers is unnecessary overhead. Accordingly, CoAP employs unreliable, asynchronous UDP protocol. UDP is a connectionless protocol and messages are sent directly to an IP address, without confirmation of successful reception. The response is transferred independently, also directly to an IP address.

Although, UDP does not provide reliability on the transport layer, CoAP implements lightweight reliability mechanisms based on confirmation of message reception with an acknowledgment response. Every message contains an ID number for duplicate detection but also, for reliability. If no ACK message was received in some time, the message is retransmitted. RESPONSE_TIMEOUT and MAX_RETRANSMIT determine how to retransmit.

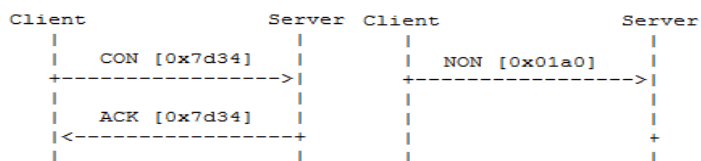


Figure 3. Reliable message delivery and unreliable message delivery

B. Request/Response Semantics

CoAP and HTTP have a similar request/response model: a CoAP client sends one or more CoAP requests to the server, which services the requests by sending CoAP responses. The difference is that the response is not sent over a previously established connection, but exchanged asynchronously over CoAP messages.

The CoAP supports the basic methods GET, POST, PUT and DELETE, which are easily mapped to the HTTP. The GET method (retrieves a representation of a resource) is a safe (it

means that can be only retrieval) and idempotent (can be invoked it multiple times with the same effects). PUT (requests that the resource on the server be updated or created) and DELETE (requests that the resource be deleted) are also idempotent. This is not the case with the POST method (requests that the representation in the payload of the request be processed) because it depends from server and it usually results in a new resource being created or the target resource being updated.

After receiving a request, a server responds with a CoAP response. A response is identified by the Code field in the header. The CoAP Response Code indicates the result of the attempt to understand the request. The responses are defined in CoAP Response Code Registry.

C. Freshness and validation model

Freshness model – To determine how fresh is the information, Max-Age Option is used. This Option can be defined by the server and the default value is 60 seconds.

Validation model - When an end-point has one or more stored responses for a GET request, but cannot use any of them (e.g., because they are not fresh), it can use the ETag Option in the GET request to give the server an opportunity to both select a stored response to be used, and to update its freshness.

D. Proxy

A proxy is a CoAP end-point that can be tasked by CoAP clients to perform requests on their behalf. This may be useful, for example, when the Proxy is caching information in order to reduce response time or when mapping CoAP to HTTP and vice versa.

A CoAP CON or NON requests, that goes over a Proxy, has a Proxy-Uri Option (that is an absolute URI). A Proxy-Uri Option defines the target server. When Proxy receive a message with this option, it retransmits it to a different Proxy or the end server defined by the Proxy-Uri.

E. HTTP-CoAP Mapping

CoAP supports a limited subset of a HTTP functionality, and thus a mapping to the HTTP is straightforward.

The CoAP request on a HTTP resource is the same as on the CoAP resource. Mapping takes place on the Proxy (whether CoAP-HTTP or HTTP-CoAP). If a CoAP GET request with a HTTP URI, as a Proxy-Uri Option, is sent to the Proxy, it is mapped to the HTTP and transmitted to the HTTP server. HTTP server reponses to the Proxy, where, HTTP-CoAP mapping takes place and message gets back to the node. It is similar with the POST, PUT or DELETE methods.

HTTP-CoAP mapping is not that straightforward, as CoAP does not support all HTTP features. Basics methods GET, POST, PUT or DELETE, are easy to translate, HEAD is possible to map on Proxy. OPTION, TRACE and CONNECT methods are not supported.

III. HARDWARE FOR EVALUATION

Telit GSM-GPRS modems are different from a typical application that uses a microcontroller for managing I/O pins on the module through the AT command interface, because they are extended with EASY Script Extension. It enables a customer program, written in a high-level open-sourced language, to run on the Telit modules [6].

In order to eliminate external controller and further simplify the programming of the sequence of operations, Telit modules introduce virtual machine and Python interpreter where is allowed to program even on high level object oriented language.

Python module has a 3MB of memory for scripts and 1,2MB of RAM memory. It uses a Python script interpreter engine v. 1.5.2+ [7].

Python scripts are simple text files stored in the NVM (Non-volatile memory) inside the Telit module. In one moment only one Python script can be executed on PY modules.

The Python script is executed in a task with the lowest priority on the Telit module, so its execution do not interfere with GSM/GPRS normal operations. Furthermore, this allows serial ports, protocol stack, etc, to run independently from the Python script. The Python script interacts with the Telit module functionalities through several built-in interfaces (Figure 4).

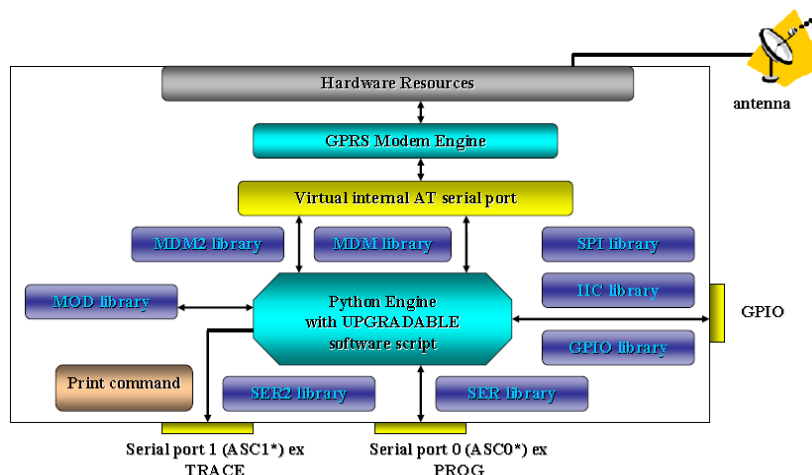


Figure 4. A Python interfaces

MDM and MDM2 is the most important interface that allows Python script to send and receive data from the network during connections [8].

IV. IMPLEMENTATION OF COAP PROTOCOL IN M2M SYSTEM

The measurement devices (resources) are installed on top and inside of the public vehicles. Devices deployed on a top of the vehicles are equipped with the gas sensors (CO, CO₂, NO₂), weather sensors (temperature, air pressure, humidity), location (GPS) and a mobile network interface (GPRS).

The GPS tracker devices are Telit GM862-GPS Modem. This device have a GPS and a GSM interface, a Board To Board connector and the SIM Card Reader. The CoAP protocol is implemented on application layer of M2M system that uses this modem.

Communicational scheme and data flow, used in this implementation are shown in Figure 5.

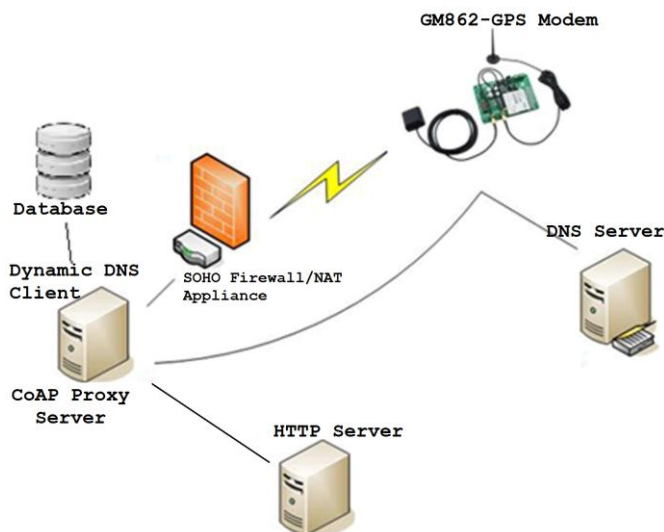


Figure 5. Implementation of CoAP protocol on GM862-GPS modem

To enable sensor devices to use CoAP as an application layer protocol, it is not enough just to send CoAP like message, the end point have to be able to process received message and to turn back response. CoAP server is was made to read message, take the payload and store it, and turn back acknowledgement message to sender. All of this allows other web applications to use data from sensors through HTTP not knowing how that information came (black box).

Server side of the system is defined in Java programming language. Access to a server is enabled through dynamic dns server. Access to dns leads to the router, which is set to forward all incoming data that are coming on the port 5683, to the predefined server in the LAN.

Modem at first opens the UDP socket and sends the request to dns. Router gets the request and forwards it to a server. The Server is able to receive and handle the request. It separates the payload from the request and put it in the Access database and external file. Afterwards, payload is inserted into a HTTP message and forwarded to a Web server.

Client side of the M2M system

Python script language was used in order to define the client side of the M2M system.

After the basic configuration is set, socket needs to be configured (connection id, package size, time to live, max reception time):

```
MDM.send('AT#SCFG=1,1,300,7,1024,50\r',0)
```

Then, UDP socket has been set (port and address):

```
MDM.send('AT#SKTSET=1,5683,"coap.dyndns-server.com"\r',0)
```

Afterwards, UDP socket should be opened with the following command:

```
MDM.send('AT#SKTOP\r',0)
```


The response, for this command, is 'CONNECT'. When this response is received, everything is ready for sending a message.

A CoAP message that needs to be sent, has header, as defined above in section 2. and a payload. Header is defined as follow: Version: 1, type NON (because of the nature of the message, information is valid only in one moment and it is not useful to send it later), Option Count: 2, Code: POST and MessageID: random unique number.

```
HEADER = '1 1 2 2 RANDOM_NUMBER
```

Option Count is defined to be two, so two options has to be sent. The first option defines Content Type of the payload (content is in a XML format), and the second indicates the Proxy-Uri. This option helps the Proxy server to reveal what is the final destination for the message. The Options are defined as CoAP specification requires.

```
OPTIONS = '1 2 \r\n41\r\n2 55\r\n'+SRV_ADDR + WEB_PATH_REP + ' ' + SRV_PORT + '\r\n'
```

Payload is sent in the following format:

```
XML_HEAD = '<EC><n>GPSData</n>'
```

```
XML_ID = '<ei>'+imei_data+'</ei><si>'+imsi_data+'</si>'
```

```
XML_GPS = '<d><gD>'+gps_data+'</gD>'
```

```
XML_TAIL = '</d></EC>\r\n\r\n'
```

Complete payload (XML) looks like this:

```
XML = XML_HEAD + XML_ID + XML_GPS + XML_TAIL
```

Server side of the M2M system

A message defined in the modem is delivered via UDP to the server-side, where the UDP socket is open and waits for message. The server side is created in Java and is waiting information on port '5683':

```
DatagramSocket serverSocket = new DatagramSocket(5683);
serverSocket.receive(receivePacket);
```

After message is received, server process it. The payload is packed into the HTTP message and that message is forwarded to the local web server. The web server, defined in a PHP, is doing only echo of the payload. This shows that it is possible to implement CoAP without making any changes to the existing installation, i.e. it is possible to forward a message to an existing HTTP server on a system platform. Java server serves as a proxy that translates CoAP messages to the HTTP and vice versa.

In addition, payload sent by the modem is stored in the local database defined in Access.

V. EVALUATION OF COAP DEPLOYMENT IN M2M SYSTEM

The evaluation of the CoAP is given as comparison of resource efficiency against previously deployed HTTP protocol for the same M2M scenario. Bandwidth and transmission speed were taken as an evaluation criteria to estimate the possible advantages of CoAP protocol in real life system.

CoAP utilizes UDP as a transport layer protocol, therefore some speed benefits can be expected on its side. These advantages are: size of the transferred data, because TCP has a

bigger segment, and data transfer speed, because the TCP protocol is a connective type, and the UDP is not.

Evaluation methodology is based on measuring the following for both protocols:

1. Transmission bandwidth
2. Transmission speed

Transmission bandwidth

For size determination Wireshark was used. Size of the payload (XML) is 163 bytes. CoAP message has a 10 bytes header and a 51 bytes options. HTTP's header is 20 bytes. Size of the whole UDP package, which is carrying a CoAP message, is 266 bytes, while the TCP segment carries 324 bytes (Table I).

TABLE I
SIZE CONSIDIRATION

	<i>CoAP [bytes]</i>	<i>HTTP [bytes]</i>
Header	10	20
Options	51	/
Payload	163	163
Whole package	266	324

Transmission speed comparison

In order to calculate a transmission time for both protocols, the following approach was used. The measurement was performed on the server side with a Wireshark, because a modem is unable to measure time smaller than a second. For the transfer of the HTTP message, a duration of the TCP connection was measured, because that represents the time needed for one message to be transferred.

On the server side, there is no information when the communication has started, i.e. when the modem has sent the first SYN message in a TCP 3-way handshake, so that has to be calculated. In the 3-way handshake procedure, we can see in the Wireshark the time when the server responded to the modem with the second ACK message and when the third message sent from the modem is received. If this time is divide with two, we get the time needed the first message to arrive from the modem to the server. In order to get more accurate time, 50 measurements were taken (Table II).

TABLE II
 . TIME NEEDED FOR THE FIRST SYN MESSAGE TO ARRIVE

Num	Time[s]	Num	Time[s]	Num	Time[s]
1	0,61815	18	0,5818	35	0,54988
2	0,79139	19	0,60003	36	0,59822
3	0,52597	20	0,60964	37	0,55408
4	0,72335	21	0,6018	38	0,57728
5	0,77444	22	0,65677	39	0,56635
6	0,63764	23	0,62256	40	0,5329
7	0,54077	24	0,56927	41	0,57866
8	0,52541	25	0,59866	42	0,59411
9	0,56662	26	0,59722	43	0,54065
10	0,53852	27	0,56072	44	0,65255
11	0,5789	28	0,6094	45	0,60033
12	0,54044	29	0,59471	46	0,53666
13	0,98407	30	0,67	47	0,60024
14	0,52432	31	0,55609	48	0,53913
15	0,58428	32	0,68542	49	0,61197
16	0,60757	33	0,56211	50	0,54632
17	0,60591	34	0,60305		

Measurement shows that average time needed to transfer the first SYN message from the modem to the server was 0.3012633 seconds. The same time has to past for message to arrived to the modem from the server when closing the connection in 2-way handshake.

Average time of the connection duration was calculated also with the 50 measurements, and also on the server side. Table III shows the measurement values.

TABLE III
TCP CONNECTION TIME

Num	Connection time [s]	Num	Connection time [s]	Num	Connection time [s]
1	8,82001	18	8,8069	35	8,73765
2	16,1032	19	8,88407	36	8,74798
3	11,0177	20	8,91578	37	9,10502
4	8,90851	21	8,81474	38	9,1503
5	16,677	22	8,79918	39	8,97075
6	8,89243	23	8,8242	40	9,28623
7	9,36813	24	8,75101	41	8,99216
8	9,19986	25	8,80949	42	8,98812
9	8,9924	26	8,96354	43	9,25768
10	9,08242	27	8,993	44	8,74795
11	8,96358	28	8,99728	45	9,05517
12	9,08824	29	9,16909	46	8,85617
13	8,79535	30	9,12084	47	8,98255
14	9,27468	31	8,99375	48	8,83881
15	8,92286	32	8,89597	49	8,91679
16	9,1812	33	9,34277	50	8,76586
17	9,11209	34	9,23638		

Average time, taken from this values is 9.322336 seconds.

Time needed to transfer HTTP message, through the TCP connection, from the modem to the server and back, is 9.9248626 seconds.

The CoAP protocol uses UDP on a transport layer so the time has to be measured using different method. The UDP is connectionless, so it does not communicate with the server before the message is sent. It sends the message on the specific address and port.

Measurements are also taken on the server side, so that the values can be compared. In the Table IV. differences between sending time and receiving time are provided. Sending time was taken from the server so that this two times can be compared.

TABLE IV
ARRIVAL OF THE COAP MESSAGE TO THE SERVER

Num	Time [s]	Num	Time [s]	Num	Time [s]
1	1,9131	18	1,93061	35	2,02831
2	1,91	19	2,17304	36	1,99722
3	1,87257	20	2,00974	37	2,00322
4	2,00883	21	1,91206	38	1,92955
5	2,10376	22	1,61611	39	1,94918
6	2,08108	23	1,96319	40	1,97978
7	1,24993	24	1,97849	41	2,13744
8	2,08792	25	1,85537	42	1,22525
9	1,93593	26	2,13685	43	1,83944
10	2,06519	27	2,03583	44	1,8629
11	2,12149	28	2,09912	45	1,92767
12	1,79254	29	1,8813	46	1,9556
13	1,78781	30	1,49717	47	1,95842
14	1,99185	31	1,94055	48	1,82597
15	1,76863	32	1,9839	49	1,7033
16	2,1393	33	1,97531	50	2,07579
17	2,30096	34	2,15925		

Average time needed for CoAP message to arrive from the modem to the server was 1.9335558 seconds. In order to compare this with HTTP, we have to take in consideration response time, so this value is multiplied with two and the value is 3.8671116 seconds.

TABLE V
AVERAGE TRANSMISSION TIME

	<i>CoAP</i>	<i>HTTP</i>
Average transmission time [s]	3.8671	9.9248

Results shows that the time for message transport is 3 times smaller if CoAP protocol is used.

In the mobile Environment System, devices are sending data to the central server approximately every 15s. As each packet is 324 bytes large, it means that one device generates and transfers 1 822,5 KB during a day. After HTTP replacement with CoAP, the total amount of bytes per device per day is reduced to 1 496,25 bytes which represents a significant reduction resulting in shorter transfer times and reduced cost of running the system.

VI. CONCLUSION

In this paper, an implementation of the Costrained Application Protocol in an Environmental Monitoring System is presented. The previously used HTTP protocol is replaced with the CoAP protocol and the performance of two solutions is evaluated in term of resource utilization. The CoAP represents the protocol of choice for the constrained networks. This protocol supports the majority of the concepts that enabled the existing Web to run, while the complexity is reduced and adjusted for capabilities of the devices with constrained resources. The time needed to transfer a CoAP message over mobile network is almost three times shorter than the time required when HTTP messages are used. Also, the size of a HTTP message is 1.2 times bigger than the CoAP message which has implications on the processing of such messages on the constrained networks side. The implementation evaluation presented in this paper shows that the CoAP protocol is easy to deploy, saves bandwidth, memory and device power in the M2M communication. Problems with reliability of the UDP are resolved with retransmission mechanism that enables devices to resend message if no response is received. For the future deployment implementation of the CoRE Linked Format as a payload format should be considered. The idea is to replace heavy XML with some lighter message format and in that way to reduce whole message on the application layer. To the constrained devices this could rapidly lower a resource consumption.

REFERENCES

- [1] CoRE Status Page, Constrained RESTful Environments (Active WG) <http://tools.ietf.org/wg/core/>
- [2] Zach Shelby, Carsten Bormann, 6LoWPAN The Wireless Embedded Internet, Wiley Series in Communication Networking and Distributed Computing
- [3] Constrained Application Protocol (CoAP) draft-ietf-core-coap-07 <http://tools.ietf.org/html/draft-ietf-core-coap-07>
- [4] Srdjan Krco, Jelena Vuckovic, and Stevan Jokic: ecoBus - Mobile Environment Monitoring, Proc. of ServiceWave 2010, Vol. 6481Springer (2010), p. 189-190, Ghent, Belgium
- [5] Koojana Kuladinithi, Olaf Bergmann, Thomas Pötsch, Markus Becker, Carmelita Görg, Implementation of CoAP and its Application in Transport Logistics
- [6] Telit, wireless solutions (2012) <http://www.telit.com>
- [7] Telit, AT GM862 Family Hardware User Guide, available at: <http://www.telit.org>, 2011.
- [8] Telit, Easy Script in Python, available at: <http://www.telit.org>, 2010.

Empirical Findings of the Awareness and Application of Barcodes in Austria

Iris Uitz*, Michael Harnisch**, Bernd M. Zunk*

* Graz University of Technology / Institute of Business Economics and Industrial Sociology, Graz, Austria

** University of Graz / Institute of Information Science and Information Systems, evolaris next level GmbH, Graz, Austria

Abstract—In mobile marketing campaigns, information can be quickly and easily transferred to consumers and thus ensure the competitive edge. Therefore, barcodes are currently a hot marketing trend. In addition to the initial use of the product labelling in warehousing and logistics, barcodes are now prevailing in virtually all areas of life due to the wide range of applications for codes. The applications vary from routing to a website, YouTube or Facebook, to making calls or paying for a product and service instead of cash or credit card. Also the area where barcodes are placed is very broad and ranges from print media, billboards or business cards to packaging. Since many different barcodes exist, companies mainly select a specific type of code. First, this decreases the amount of data that has to be encoded and, second, it improves the awareness and recognition factor of the symbol. So far, only limited research has been conducted to show the consumers awareness and application of different barcodes. On the basis of an online survey, this paper analyses differences in attitude and awareness of multiple codes. Furthermore, reasons why people scan codes or why they refuse these new marketing channels are discussed.

Keywords—Barcode, Two-Dimensional Code, QR Code, HCCB, Mobile Marketing

I. INTRODUCTION

Barcodes are a modern type of illustration and give communication a whole new approach. With the rapid development of smartphones and the booming mobile internet applications that come along with it, barcodes are being used more frequently and new barcodes originate.

In 2011 472 million smartphones were sold worldwide and 17% of the global population already has an active mobile-broadband subscription [1]. Since the proliferation of smartphones will increase significantly in the coming years (see Figure 1 **Error! Reference source not found.**), also the use of mobile internet and mobile services increases [2], [3].

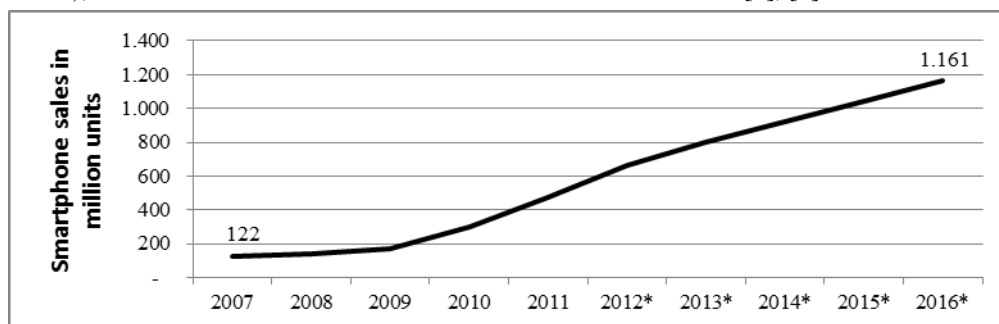


Fig. 1 Smartphone Sales Worldwide from 2007 until 2016, in Million Units [2], [4]

Due to this increasing spread, the consumer behavior as well as the pattern of web usage is changing. Communication and shopping is just a 'click away' and available anytime and anywhere. In addition to the already improved usability of mobile services the biggest obstacle to more widespread use of mobile internet is the media break between the real world and the content of the Internet. The big challenge is to provide customers an easy access to information and services. A very simple solution is the technology of optical coding using barcodes. In this way, mobile smartphone users are able to capture all the information they need easily – and interact with this information immediately – by taking a picture of the barcode with their smartphone as they walk by the advertisement and let the application do the rest.

In addition to the initial use of the product labeling in warehousing and logistics, barcodes are now prevailing in virtually all areas of life due to the wide range of applications for codes. The applications vary from routing to a website, YouTube or Facebook in order to provide detailed information about the product and tracking packages and mail, to making calls or paying for a product or service instead of cash or credit card [5]. Also the application area is very broad. Barcodes are now found on nearly every item: on electronic parts, packaging, newspapers, billboards, business cards or even coffee mugs or T-Shirts.

Since a great diversity of barcodes exist, for marketers, understanding which consumer segments scan [...] codes, the source and location of these scans, and the resulting information delivered, is crucial [...].’ (Mark Donovan, [6])







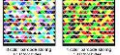
This paper aims to reveal motives of consumers for scanning barcodes as well as differences in awareness, attitude and appliance of different codes.

II. BARCODES

The technology of optical coding offers an enormous number of different code combinations. The best known and most frequently used ones are encountered in barcodes. The initial intention of barcodes was in the rationalization of warehousing. The information in so called 1D codes such as EAN and UPS, is encoded in one-dimensional horizontal bars. As a consequence only a limited number and type of data can be stored. Due to the restricted space, these codes reached their limitations in certain areas of application [7]. With the idea of 'stacking' multiple 1D codes above each other, so called stack codes, the first principle of 2D codes emerged. Each code type, such as staple codes or matrix codes, contains many different versions that are optimized in terms of data capacity, a quick readability or robustness against reading error for a specific purpose. Today's developments represent an increase in data capacity by the distinction of different colors (3D codes) or the chronology of code sequences (4D codes).

Above all, the ever-growing smartphone penetration accelerates application development and creates new barcodes every day. Currently there are over 40 different 2D and 3D barcodes on the market [8]. Table I gives an overview of selected barcodes.

TABLE I
OVERVIEW OF DIFFERENT BARCODES

		Developer (Country) Year	Capacity	Main field of application	Standards
EAN Code		Developed out of UPC Code by RCA (USA) In Europe 1976	e.g. PDF 14 (2D Code): Numeric: 2.710 Alphanumeric: 1.850 Binary: 1.018	Office Automation	ISO, EAN International, BSI, DIN, NEN etc.
QR Code		Denso (Japan) 1994	Numeric: 47.089 Alphanumeric: 2.953 Binary: 1.817	Logistics, Factory Automation, Mobile phones, Advertisement	AIM, JIS, ISO
Shotcode		University Cambridge (UK) 1999	Consists of 40 bit pro data	Advertisement	
BeeTagg		Convision AG (Switzerland) 2007	10 ³⁴ different BeeTaggs	No directly applicable information, Mobile Tagging	
DataMatrix / Semacode		RVSI Acuity CiMatrix (USA) 2006	Numeric: 3.116 Alphanumeric: 2.355 Binary: 1.556	Identifying parts in automation, pharma industry, document handling	AIM, ISO
Aztec Code		Welch Allyn (USA) 1995	Numeric: 3.832 Alphanumeric: 30.67 Binary: 1.914	Eticket, bills	AIM, ISO
Microsoft Tag High Capacity color barcode (HCCB)		Microsoft (USA) 2007	1000 per square inch	Biometric ID, Labeling, Advertisement	ISAN-IA

According to its area of operation, different demands are made on the various codes [9]. These include data capacity, which is referred to the maximum amount of data, as well as the data density, which indicates the concentration of the data related to the required area. Another important feature is the error detection and correction, which increases data security, because also dirty, poorly printed or partly damaged codes become readable. The robustness of a code is also essential. It takes the tolerance of rotation, strain, different angles, lighting conditions or damage into account. Aesthetic and conspicuousness of a code play a significant role for the acceptance by end users

III. METHODOLOGY AND SAMPLE

This study employs a survey method in form of a quantitative analysis. The survey was conducted in the first quarter of 2012 by using an online questionnaire via the survey software tool 'limesurvey'.

The survey consisted of 79 questions, clustered into 13 sections. Section one had the aim to query the smartphone usage. After a section relating to Social Media behavior, the barcode section (section three) contained all relevant questions regarding the awareness, scanning behavior and preference of barcodes. Each further section queried the use of one specific barcode, based on the knowledge in section three. Hence, the section was skipped if the specific code was not known. In the end demographic data was obtained.

The main target group of this survey was Austria and especially the younger generation. A random sampling technique was adopted for questionnaire distribution. The questionnaire was sent out to schools, universities and associations. To reduce biases in format and content as well as to enhance validity of the questionnaire a pretest of 25 people was conducted prior to the actual data collection.

In total more than 1,059 participants were interviewed. After considering unanswered and not valid forms 784 questionnaires build the basis for the analysis. The demographic profile of the respondents is shown in Table II.

TABLE II
DEMOGRAPHIC OVERVIEW OF THE TESTED SAMPLE POPULATION

Item		Frequency	%
Gender	Male	278	35.5
	Female	506	64.5
Age	Average Age		26.78 years
	Age <30	655	83.5
	30-50	100	12.8
Smartphone User	>50	29	3.7
	Yes	463	59.1
	No	317	40.4
	Don't know	4	0.5

IV. RESULTS

A. General results

The smartphone penetration is generally increasing in Austria. From the second quarter of 2011 to the second quarter of 2012, the penetration of smartphones in Austria increased from 43% to 47% of all mobile phone users [10]. Also our study reflected this tendency, by concluding a market share of 59.1% of smartphones within all respondents. The higher penetration rate is explainable by the average age of the sample group. Smartphone penetration is generally much higher in younger sample groups than in older sample groups. The Austrian Internet Monitor found, that in target groups ranging from ages 14 to 29, smartphone penetration rates can be as high as 72% of all mobile phone users. (Integral, 2012) Surprisingly, our study revealed that males are more likely to have a smartphone (66.2%) than females (55.1%).

As Figure 2 (right illustration) shows, most smartphone users possess a mobile device that runs on Android (58.35%). Subsequently the iPhone establishes itself clearly as number two in the market (21.34%). The results reflect the current situation on the smartphone sector (see Figure 2 - left).

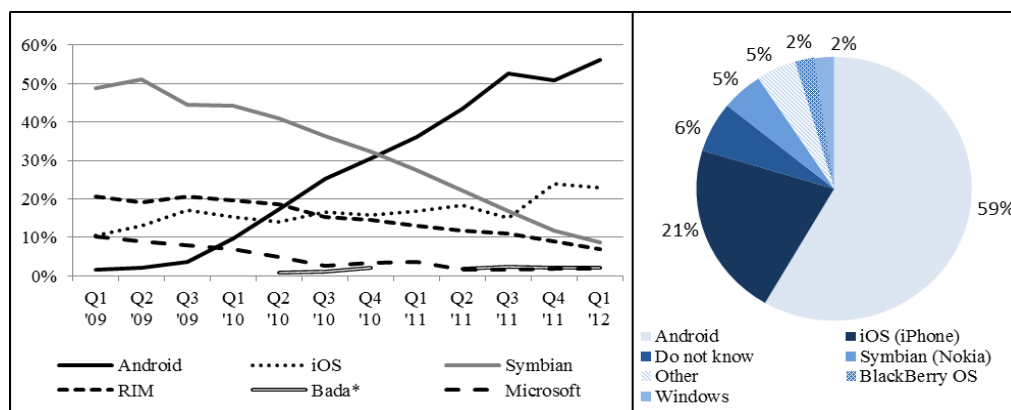


Fig. 2 left: Market Share of Leading Operating Systems on Sales of Smartphones Worldwide [Gartner via 11] – right: Results of Survey: Operating Systems of Smartphones

Our survey also indicates that the average respondent already possesses his phone for about a year, for most (67.10%) it is the first smartphone.

B. Barcode results

The survey depicted that 35.6% of the respondents are able to scan a code with their smartphone. However, 20.0% are not aware whether their smartphone is capable of scanning a code. 56.99% of the participants who are able to scan a code already scanned a barcode at least once. On average the selected barcodes have been scanned 13.41 times (by people who have ever scanned a barcode), with a deviation of 10.85, and mainly for private reasons (80.57%). Interestingly, on average female participants did not scan as many barcodes as male participants (see Table III).

TABLE III

MEAN VALUE OF SCANNED BARCODES BY GENDER

Gender	Mean	N	Std. Deviation
Female	10.71	76	9.920
Male	15.67	91	11.004
Total	13.41	167	10.82

Reasons for not scanning a code are shown in Figure 3. The main reason lies in the lack of interest. Most respondents do not see any benefit in scanning a code. Concerns about the security of data are also an issue.

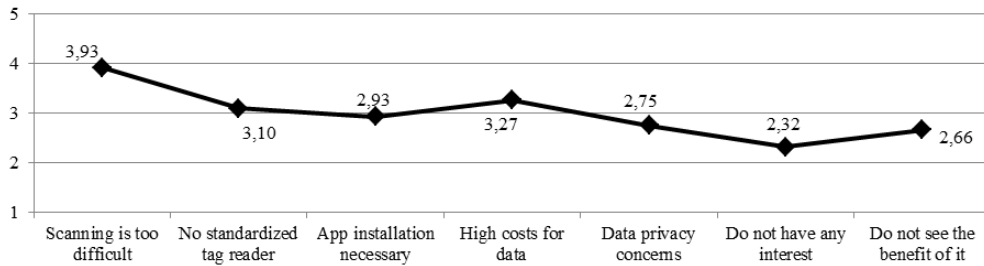


Fig. 3 Reasons for Not Scanning a Code

Grades: 5=totally agree, 4=rather agree, 3=neutral, 2=rather disagree, 1=totally disagree

All respondents were asked which barcodes they are aware of and if they use certain types of codes. The results indicate, that the best known barcode is the EAN Code (40.28%), followed by the QR Code (38.13%) and the Aztec Code (10.93%). Besides the DataMatrix Code (6.52%), all of the other requested codes have hardly been noticed by the participants. It is very interesting to see that QR Codes have caught the consumers' eyes nearly as much as EAN Codes, which are on every packaging.

The findings of this survey show that 48.09% find barcodes very useful or rather useful. Only 10.08% are of the opinion that barcodes are (rather) useless. This leaves a majority of 41.80% indifferent respectively neutral about the usefulness of. Results indicate, that users of smartphones which are running on Symbian or Android find barcodes a little more useful than users who work with an iOS- or Windows-based smartphone (see Table IV).

TABLE IV

USEFULNESS OF BARCODES BY OPERATING SYSTEM

What type of Smartphone are you using?	Mean	N	Std. Deviation
iOS (iPhone)	2.25	101	0.974
Android	2.47	257	0.976
Blackberry OS	2.37	16	1.025
Windows	2.23	13	0.927
Symbian (Nokia)	2.53	34	1.022
Do not know	2.62	26	0.804
Total	2.42	447	0.972

Additionally it was found, that the perceived usefulness of barcodes differ among the gender of participants. While male respondents on average find that barcodes are rather useful (factor = 2.19), female respondents have an attitude towards the usefulness of barcodes that is much more neutral (factor = 2.61). Also, the standard deviation of the perceived usefulness of male respondents is slightly higher than the standard deviation of the perceived usefulness of barcodes by female participants (see Table V).

TABLE V

MEAN VALUE OF SCANNED BARCODES BY GENDER			
Gender	Mean	N	Std. Deviation
Female	2,61	506	0,931
Male	2,19	278	1,022

The survey also contained an open question about considered advantages and disadvantages. Thereby the answers were clustered into different categories and the values were calculated based on the allocation. As already mentioned above, the use of barcode poses potential danger to business and consumer users of smartphones. 35.68% find barcodes risky in particular because it is uncertain to which destination they are link to. 22.11% see a disadvantage of bar codes in the necessity of decryption tools such as a scanner and a smartphone. 10.05% also feel that the reading of bar codes often does not work. Also the requisite internet connection is considered to be a disadvantage of barcodes (6.03%). Surprisingly, some people also specified that barcodes reduce social contacts because they for example replace direct contact to the customer. A possible explanation might be found in the virtual store applications of codes. If e-commerce is the main procurement form of goods the contact between fellow human beings could get lost. Also the example of Applebee's Neighborhood Grill and Bar where customers were invited to scan a code of a tabletop place card while waiting for their meal, decreases social interaction [11].

However, advantages of barcodes were seen in the easy way of gaining more information about products and services (31.82%). Even though 28.79% see the advantage of barcodes in the easy and rapid conducting of activities, some respondents stated that they do not see any time-savings through the use of barcodes (value of 7.04% of denoted disadvantages). Furthermore barcodes represent a way to automate as well as standardize business processes (14.65%). 10.61% find barcodes easy to use and 6.31% see the advantage of bar codes in the low space consumption.

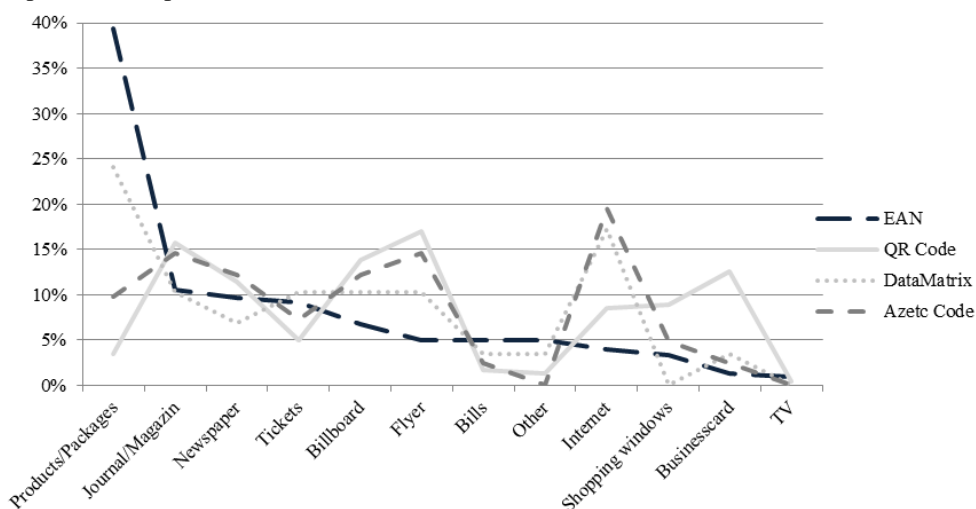


Fig. 4 Where Did You Scan Codes of?

Barcodes are mainly used for the purpose of information retrieval. Therefore codes of products and packages as well as flyers and billboards are scanned. Aztec- and DataMatrix Codes are also scanned from Internet sites whereas business cards are also a popular scanning object for QR Codes. Another interesting result is that geocaching and coupons are no motivational factors for an increased use of codes. Furthermore, the mobile payment is hardly prevalent.

Figure 5 summarizes the results regarding the awareness, appliance and optical preference of selected barcodes.

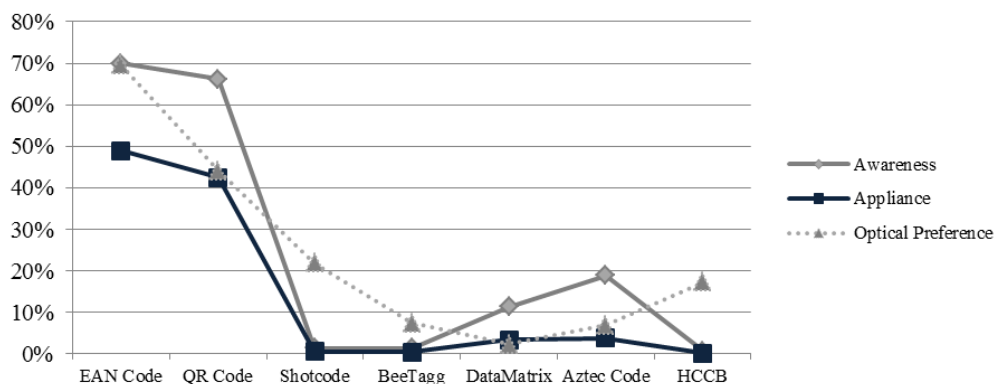


Fig. 5 Relation of Awareness, Appliance and Optical Preference of Queried Codes

V. DISCUSSION

Since the study is based on a quantitative research design, some answers could not be retrieved in full deepness. Especially necessary requirements for a more intensive use of barcodes are an interesting issue and should be investigated further. Also the reason why barcodes with a more appealing design are rather unknown should be examined in detail. All respondents came from one cultural area; cross-cultural differences cannot be considered in this research. Limitations in this study also lie in the predominantly young and female sample.

VI. CONCLUSION

Based on an online survey with 784 valid respondents, this study shows that 59.1% of the sample possess a smartphone and 35.6% are able to scan a code with it. Only 3% of the respondents have not noticed any of the queried codes. The results indicate that the EAN Code is the best-known code, followed by the QR Code. The rather trendy codes, such as Shotcode, BeeTagg or HCCB, are almost unknown. The appliance of the codes reflects the same results. Even though barcodes are considered to be useful, the scanning habit is not widely accepted. EAN Codes have been scanned 4.56 times on average by people who know it, QR Codes 2.92 times and the other codes less than once. The main use lies in receiving information. The use of scanning barcodes in order to get discounts or pay by mobile is not widely spread in our

sample. Because of the simple and already inured design the EAN Code is also the code that appeals most to the respondents.

Codes will become more and more important in marketing management. In Japan, for example, it is already well established to scan codes in order to receive information or shop for goods. With the increasing smartphone penetration the easy and effective way of interacting with advertisement will also increase in Austria. Nevertheless, barcode readers should be preinstalled on every smartphone.

Since fun is the second motivation for scanning codes, the code's looks and the integration in billboards, flyer, journals etc. are important to attract attention and hence action. Therefore companies could use QR Codes in different, flashier colors than just black and white. And, since the error correction is pretty high at QR codes also pictures can be embedded in the code in order to increase the fun factor.

This paper gives an overview of the perception and scanning behavior of different barcodes with the aim to help struggling companies to determine the strategy of marketing campaigns.

ACKNOWLEDGMENT

This paper is an extended version of a previous published conference contribution in: Ivkovic, M./Bach, M./Simicevic, V.(Ed.): Proceedings of the IBC 2012, 1st International Internet & Business Conference, Zagreb (2012),Croatia

REFERENCES

- [1] The International Telecommunication Union (ITU), "The World in 2011: ICT Facts and Figures," Online article, available at: <http://www.itu.int/ITU-D/ict/facts/2011/index.html> (25 October 2011)
- [2] Gartner, "Gartner Says Worldwide Smartphone Sales Soared in Fourth Quarter of 2011 With 47 Percent Growth," Online article, available at: <http://www.gartner.com/it/page.jsp?id=1924314> (15 February 2012)
- [3] Ottogroup, "Go Smart 2012: Always-in-touch," Online article, available at: http://www.ottogroup.com/media/docs/de/studien/go_smart.pdf (12 February 2012)
- [4] R.L. Llamas, W. Stofega, "Worldwide Smartphone 2012–2016 Forecast and Analysis," IDC Study, 2012.
- [5] I. Uitz, M. Harnisch, "Der QR-Code – aktuelle Entwicklungen und Anwendungsbereiche," *Informatik Spektrum*, Vol. 35, Nr. 5, 2012, p. 339-347.
- [6] ComScore, Inc., "14 Million Americans Scanned QR Codes on their Mobile Phones in June 2011," Online article, available at: http://www.comscore.com/Press_Events/Press_Releases/2011/8/14_Million_Americans_Scanned_QR_or_Bar_Codes_on_their_Mobile_Phones_in_June_2011 (14 August 2011)
- [7] C. Rosol, "Das Kreuz Mit Den Strichen," Online article, available at: http://www.nzz.ch/nachrichten/hintergrund/wissenschaft/das_kreuz_mit_den_strichen_1.666225.html (12 December 2011)
- [8] Webster's Online Dictionary, "Definition of Barcode," Online article, available at: <http://www.websters-online-dictionary.org/definitions/Barcode> (19 September 2011)
- [9] M. Hegen, "Mobile Tagging: Potenziale von QR-Codes im Mobile Business," *Diplomica*, 2011, pp. 36-39.
- [10] Integral, "APP-Banking im Vormarsch - Integral Austrian Internet Monitor 2012: Smartphones & Apps", Online article, available at:

<http://www.erstegroup.com/de/Downloads/0901481b800b447d.pdf;jsessionid=Bh9GQD3WSyGT2Z2ByGWp95Yn3SZG5GfmVNnQzhnK0sHTvmzhyrBC!-1953785057> (15 June 2012)

- [11] Statista, “Prognostizierter Absatz von Smartphones,“ Online article, available at:
<http://de.statista.com/statistik/daten/studie/12865/umfrage/prognose-zum-absatz-von-smartphones-weltweit/>;
available at: <http://de.statista.com/statistik/daten/studie/12856/umfrage/absatz-von-smartphones-weltweit-seit-2007/> (6 February 2012)
- [12] L. Johnson, “Applebee's franchisee increases lunch sales by 9.8pc with QR codes,“ Online article, available at:
<http://www.mobilecommercedaily.com/2011/10/28/applebeees-increases-lunch-sales-by-9-8pc-with-qr-codes>
(30 October 2011)

Modeling and Navigation of an Autonomous Quad-Rotor Helicopter

Gyula Mester*, Aleksandar Rodic**

* University of Szeged/Faculty of Engineering, Robotics Laboratory, Szeged, Hungary

** University of Belgrade/Institute Mihajlo Pupin, Robotics Laboratory, Belgrade, Serbia

Abstract— Autonomous outdoor quad-rotor helicopters increasingly attract the attention of potential researchers. Several structures and configurations have been developed to allow 3D movements. The autonomous quad-rotor architecture has been chosen for this research for its low dimension, good maneuverability, simple mechanics and payload capability. This paper presents the modeling and navigation of an autonomous outdoor quad-rotor helicopter.

Keywords—Autonomous, modeling, navigation, quad-rotor helicopter

I. INTRODUCTION

The quad-rotor helicopter configuration is well known and has been studied since the beginning of 1900s. In 1907, the first known quad-rotor helicopter, Gyroplane No. 1 flew. Autonomous quad-rotor helicopters increasingly attract the attention of potential researchers. In fact, several industries require robots to replace men in dangerous, boring or onerous situations. A wide area of this research is dedicated to aerial platforms. Several structures and configurations have been developed to allow 3D movements [1]-[12], there are blimps, fixed-wing planes, single rotor helicopters, bird-like prototypes, quad-rotors, etc. Each of these has advantages and drawbacks. The vertical take-off and landing requirements exclude some of the aforementioned configurations. However, the platforms which show these characteristics have a unique ability for vertical, stationary and low speed flight. The electrically powered four-rotor quad-rotor helicopter architecture has been chosen for this research for its low dimension, good maneuverability, simple mechanics and payload capability (Fig. 1).



Figure 1 Quad-rotor helicopter

This structure can be attractive in several applications, in particular for surveillance, for imaging dangerous environments and for outdoor navigation and mapping.

The paper is organized as follows: Section 1: Introduction. In Section 2, the modeling of the quad-rotor helicopter is presented. In Section 3 the control strategy are presented. In Section 4, the GPS navigation of the quad-rotor helicopter is illustrated. Conclusions are given in Section 5.

II. MODELING OF THE QUAD-ROTOR HELICOPTER

The model of the quad-rotor helicopter and the rotational directions of the propellers can be seen in Fig. 2. The rotor pair 2 and 4 rotates clockwise direction and the rotor pair 1 and 3, anticlockwise direction. A quad-rotor helicopter has fixed pitch angle rotors and the rotor speeds are controlled in order to produce the desired lift forces.

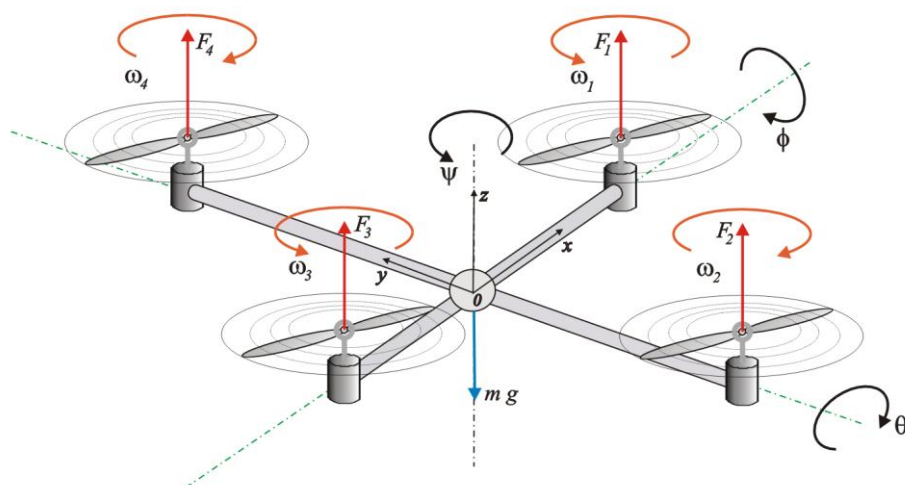


Figure 2 The model of the quadrotor helicopter

A. Actuators of the Quadrotor Helicopter

The quadrotor helicopter has four actuators - brushless DC motors which exert lift forces F_1 , F_2 , F_3 , F_4 proportional to the square of the angular velocities of the rotors. Actually, four motor driver boards are needed to amplify the power delivered to the motors. Their rotation is transmitted to the propellers which move the entire structure.

B. Sensor System of the Quadrotor Helicopter

Two types of sensors are used for measuring the robot attitude and for measuring its height from the ground. For the first, an Inertial Measurement Unit (IMU) was adopted, while the distance was estimated with a Sound Navigation And Ranging (SONAR) and an InfraRed (IR) modules. There are: accelerometers and angular velocity sensors on the board of the quad-rotor helicopter. The concept of the vision system is originated from motion-stereo approach. The

camera is attached to the quadrotor helicopter. The data processing and the control algorithm are handled in the Micro Control Unit (MCU) which provides the signals to the motors.

C. Coordinate Systems for Navigation

To describe the motion of a 6 DOF rigid body it is usual to define two reference frames [1]:

- the earth inertial frame (E-frame), and
- the body-fixed frame (B-frame), Fig. 3.

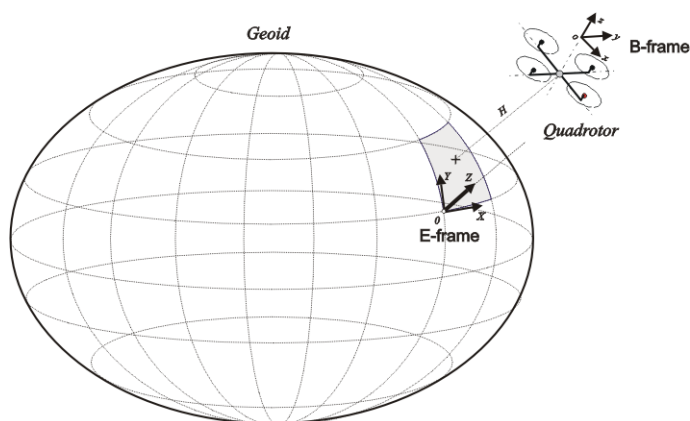


Figure 3 Earth- and Body-frame used for modeling of the quad-rotor system

The equations of motion are more conveniently formulated in the B-frame because of the following reasons:

- The inertia matrix is time-invariant.
- Advantage of body symmetry can be taken to simplify the equations.
- Measurements taken on-board are easily converted to body-fixed frame.
- Control forces are almost always given in body-fixed frame.

The E-frame ($OXYZ$) is chosen as the inertial right-hand reference. Y points toward the North, X points toward the East, Z points upwards with respect to the Earth, and O is the axis origin. This frame is used to define the linear position (in meters) and the angular position (in radians) of the quad-rotor.

The B-frame ($oxyz$) is attached to the body. x points toward the center of gravity of the quad-rotor front, y points toward the quad-rotor left, z points upwards and o is the axis origin. The origin o is chosen to coincide with the center of the quad-rotor cross structure. This reference is righthand too. The linear velocity v (m/s), the angular velocity Ω (rad/s), the forces F (N) and the torques T (Nm) are defined in this frame. The linear position of the helicopter (X, Y, Z) is determined by the coordinates of the vector between the origin of the B-frame and the origin of the E-frame according to the equation.

The angular position (or attitude) of the helicopter (Φ, θ, ψ) is defined by the orientation of the B-frame with respect to the E-frame. This is given by three consecutive rotations about the

main axes which take the E-frame into the B-frame. In this paper, the “roll-pitch-yaw” set of Euler angles were used. The vector that describes the quad-rotor position and orientation with respect to the E-frame can be written in the form:

$$s = [X \ Y \ Z \ \Phi \ \theta \ \psi]^T \quad (1)$$

The rotation matrix between the E- and B-frames has the following form:

$$\mathbf{R} = \begin{bmatrix} c_\psi c_\theta & -s_\psi c_\theta + c_\psi s_\theta s_\phi & s_\psi s_\theta + c_\psi s_\theta c_\phi \\ s_\psi c_\theta & c_\psi c_\theta + s_\psi s_\theta s_\phi & -c_\psi s_\theta + s_\psi s_\theta c_\phi \\ -s_\theta & c_\theta s_\phi & c_\theta c_\phi \end{bmatrix} \quad (2)$$

The corresponding transfer matrix has the form:

$$\mathbf{T} = \begin{bmatrix} 1 & s_\phi t_\theta & c_\phi t_\theta \\ 0 & c_\phi & -s_\phi \\ 0 & s_\phi / c_\theta & c_\phi / c_\theta \end{bmatrix} \quad (3)$$

Where c_θ and s_θ represent $\cos(\theta)$ and $\sin(\theta)$ respectively.

D. Kinematical Model of the Quad-rotor Helicopter

The system Jacobian matrix, taking (2) and (3), can be written in the form:

$$\mathbf{J} = \begin{bmatrix} \mathbf{R} & \mathbf{0}_{3 \times 3} \\ \mathbf{0}_{3 \times 3} & \mathbf{T} \end{bmatrix} \quad (4)$$

where $\mathbf{0}_{3 \times 3}$ is a zero-matrix. The generalized quad-rotor velocity in the B-frame has a form [2]:

$$\mathbf{v} = [\dot{x} \ \dot{y} \ \dot{z} \ \dot{\phi} \ \dot{\theta} \ \dot{\psi}]^T \quad (5)$$

Finally, the kinematical model of the quad-rotor helicopter can be defined in the following way:

$$\dot{\mathbf{s}} = \mathbf{J} \cdot \mathbf{v} \quad (6)$$

E. Dynamic Model of the Quad-Rotor Helicopter

Dynamic modelling of the quadrotor helicopter is a well elaborated field of aeronautics. The dynamics of a generic 6 DOF rigid-body system takes into account the mass of the body m and its inertia matrix \mathbf{I} .

Two assumptions have been done in this approach:

- The first one states that the origin of the body-fixed frame is coincident with the center of

mass (COM) of the body. Otherwise, another point (COM) should be taken into account, which could make the body equations considerably more complicated without significantly improving model accuracy.

- The second one specifies that the axes of the B-frame coincide with the body principal axes of inertia. In this case the inertia matrix I is diagonal and, once again, the body equations become simpler.

Each rotor produces moments as well as vertical forces. These moments were observed experimentally to be linearly dependent on the forces at low speeds. There are four input forces and six output states $(x, y, z, \psi, \theta, \phi)$ and, therefore the quad-rotor is an under-actuated system. The rotation direction of two of the rotors are clockwise while the other two are counter clockwise, in order to balance the moments and to produce yaw motions as needed. The equations of motion can be written using the force and moment balance, yielding:

$$\ddot{x} = \frac{\left(\sum_{i=1}^4 F_i \right) (c_\phi s_\theta c_\psi + s_\phi s_\psi) - K_x \dot{x}}{m} \quad (7)$$

$$\ddot{y} = \frac{\left(\sum_{i=1}^4 F_i \right) (s_\phi s_\theta c_\psi + c_\phi s_\psi) - K_y \dot{y}}{m} \quad (8)$$

$$\ddot{z} = \frac{\left(\sum_{i=1}^4 F_i \right) (c_\phi c_\psi) - K_z \dot{z} - G}{m} \quad (9)$$

$$\ddot{\psi} = l \cdot \frac{(-F_1 + F_2 + F_3 - F_4 - K_\psi \dot{\psi})}{J_x} \quad (10)$$

$$\ddot{\theta} = l \cdot \frac{(-F_1 - F_2 + F_3 + F_4 - K_\theta \dot{\theta})}{J_y} \quad (11)$$

$$\ddot{\phi} = \frac{(-M_1 + M_2 + M_3 - M_4 - K_\phi \dot{\phi})}{J_z} \quad (12)$$

The factors K_j in (7)-(12) given above are the air resistance coefficients to be determined experimentally. J_x, J_y, J_z are the moments of inertia with respect to the particular axes.

III. MODELING OF THE CONTROL STRATEGY

Together with modeling, the determination of the control algorithm structure is very important for improving stabilization. Controlling a autonomous quad-rotor helicopter is

basically dealing with highly unstable dynamics and strong axes coupling. In addition to this, any additional on-board sensor increases the autonomous quad-rotor helicopter total weight and therefore decreases its operation time. The control system of the autonomous quad-rotor helicopter requires accurate position and orientation information [5], [6], [8] [9] [10]. In this section we present a control strategy to stabilize of the quad-rotor. Fig. 4 shows the block diagram of the quad-rotor control system.

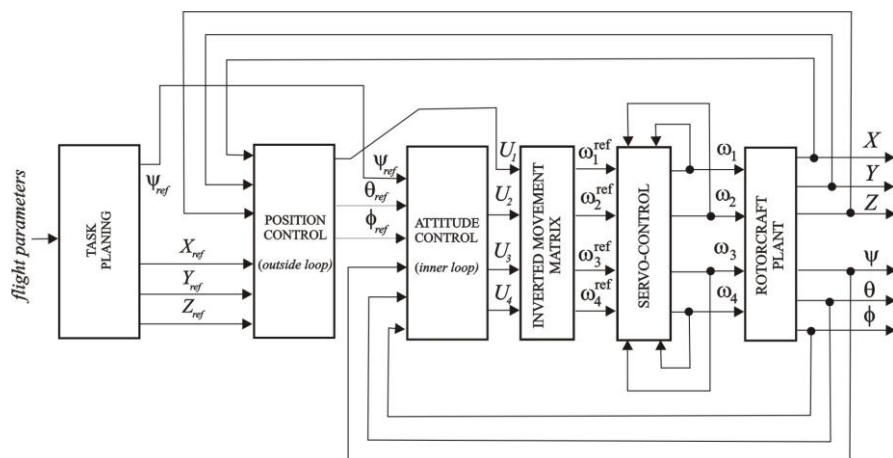


Figure 4 The block diagram of the quad-rotor helicopter control system

The task planning block is in debt to determine desired, i.e. referent 3D rotorcraft trajectory of flight as well as to propose the referent flight speed along the trajectory. The task planning block generates referent path based on flight parameters and quad-rotor task imposed.

Position control block has to ensure accurate 3D trajectory tracking. It represents so called outside control loop. Based on sensory information (GPS, IR, SONAR) about the referent positions (speeds) and corresponding actual ones defined in the inertial coordinate system (E-frame), the position controller calculates referent attitude position of quad-rotor body (pitch and roll angle) that have to enable desired motion.

Inner control block represents the core of the control scheme. It is responsible for the attitude control of quad-rotor system. Appropriate attitude control ensures in an indirect way required flight performances in the particular directions of motion such as longitudinal, lateral as well as vertical. Inner control block processes the task and sensor data and provides a signal for basic movements which balances the position error. The essence of building control scheme presented in Fig. 4 is that by controlling a body attitude (within an inner loop) it is enabled controlling of the rotorcraft movements in the coordinate directions co-linear with the axes of the inertial system

Inverted Movements Matrix block is used to compute the propeller's squared speed from the four basic movement signals.

Variety of control algorithms can be implemented within the flight controller presented in Fig. 4. These are: (i) conventional PID regulator, (ii) backstepping method and (iii) knowledge-based Fuzzy Inference System (FIS).

IV. GPS NAVIGATION OF THE AUTONOMOUS QUAD-ROTOR HELICOPTER

The trajectory of the autonomous quad-rotor can be introduced by GPS coordinates (e.g. $P_{GPS}(j)$) as shown in Fig. 5. The autonomous quad-rotor helicopter is requested to track the imposed trajectory between the particular points ($j=1, \dots, n$) with satisfactory precision, keeping the desired attitude and height of flight [11], [12]. The autonomous quad-rotor helicopter checks for the current position: X and Y by use of a GPS sensor and/or electronic compass. Also, the altitude is measured by a barometric sensor. An on-board microcontroller calculates the actual position deviation from the imposed trajectory given by successive GPS positions $P_{GPS}(j)$. It localizes itself with respect to the nearest trajectory segment, by calculation of the distances: δ_1 or δ_2 .

Gyroscopes provide angular velocity measurements with respect to inertial space. With recent developments in gyroscope technology, their usage in various fields is observably increasing. In combination with accelerometers, gyroscopes are used in position, velocity, and attitude computation in a variety of navigation and motion tracking applications for aircraft and robots [14-15]. By providing angular velocity measurements, gyroscopes can also be used in angular orientation estimation.

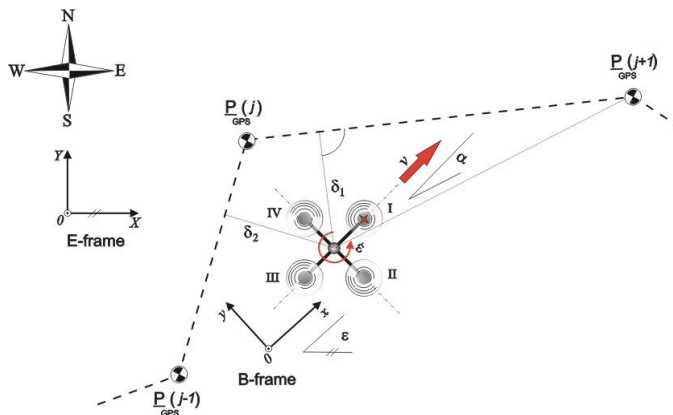


Figure 5 The block diagram of the quad-rotor helicopter control system

Using the gyroscope, the autonomous quad-rotor helicopter determines desired azimuth of flight α (Figure 5) and keeps the desired direction of flight. The height of flight is also controlled to enable the performance of the imposed mission (task).

V. CONCLUSIONS

We presented the modeling and navigation of an autonomous quad-rotor helicopter in a outdoor scenario. The main aspects of modeling of rotorcraft kinematics and rigid body dynamics, spatial system localization and navigation of autonomous quad-rotor helicopter in outdoor scenario are considered in the paper. The control strategy is presented. The GPS navigation of the autonomous quad-rotor helicopter is illustrated.

ACKNOWLEDGMENT

This work was supported by the innovation project ‘Research and Development of Ambientally Intelligent Service Robots’, TR-35003, 2011-2014, funded by the Ministry of Science of the Republic Serbia and partially supported by the TÁMOP-4.2.2/08/1/2008-0008 program of the Hungarian National Development Agency.

REFERENCES

- [1] Aleksandar Rodic, Gyula Mester, Ivan Stojković, “Qualitative Evaluation of Flight Controller Performances for Autonomous Quadrotors”, in *Intelligent Systems: Models and Applications*, Endre Pap (Ed.), Topics in Intelligent Engineering and Informatics, Vol. 3, Part. 2, ISSN 2193-9411, e-ISSN 2193-942X, ISBN 978-3-642-33958-5, e-ISSN 978-3-642-33959-2, DOI 10.1007/978-3-642-33959-2_7, Springer, 2012, pp. 115-134.
- [2] Aleksandar Rodic, Gyula Mester, *The Modeling and Simulation of an Autonomous Quad-Rotor Microcopter in a Virtual Outdoor Scenario*, *Acta Polytechnica Hungarica, Journal of Applied Sciences*, Vol. 8, Issue No. 4, Budapest, Hungary, 2011, pp. 107-122.
- [3] Aleksandar Rodic, Gyula Mester, "Ambientally Aware Bi-Functional Ground-Aerial Robot-Sensor Networked System for Remote Environmental Surveillance and Monitoring Tasks", *Proceedings of the 55th ETRAN Conference, Section Robotics*, Vol. RO2.5, pp 1-4, ISBN 978-86-80509-66-2, Banja Vrućica, Bosnia and Herzegovina, 2011.
- [4] Aleksandar Rodic, Gyula Mester, "Modeling and Simulation of Quad-rotor Dynamics and Spatial Navigation", *Proceedings of the SISY 2011, 9th IEEE International Symposium on Intelligent Systems and Informatics*, pp 23-28, ISBN: 978-1-4577-1973-8, DOI: 10.1109/SISY.2011.6034325, Subotica, Serbia, 2011.
- [5] C. Lebres, V. Santos, N. M. Fonseca Ferreira and J. A. Tenreiro Machado, “Application of Fractional Controllers for Quad Rotor, Nonlinear Science and Complexity”, Part 6, DOI: 10.1007/978-90-481-9884-9_35, Springer, 2011, pp. 303-309.
- [6] J. Coelho, R. Neto, C. Lebres, V. Santos: “Application of Fractional Algorithms in Control of a Quad Rotor Flight”, *Proceedings of the 2nd Conference on Nonlinear Science and Complexity*, Porto, Portugal, July 28-31, 2008, pp. 1-12.
- [7] Tommaso Bresciani, Modelling, “Identification and Control of a Quadrotor Helicopter”, Department of Automatic Control, Lund University, ISSN 0280-5316, ISRN LUTFD2/TFRT/5823.SE, October 2008.
- [8] B. Siciliano, O. Khatib, Eds., *Handbook of Robotics*, Springer, ISBN: 978-3-540-23957-4, 2008, pp. 391-410.
- [9] Barnes W., McCormick, W., *Aerodynamics Aeronautics and Flight Mechanics*. New York: Wiley, 1995.
- [10] Gordon Leishman, J., *Principles of Helicopter Aerodynamics*, Second Edition, Cambridge University Press, 1995.
- [11] Etkin, B., Reid L. R., *Dynamics of Flight- Stability and Control*. JohnWiley & Sons. New York, 1996.

- [12] Castillo, P. Dzul, A. Lozano, R. Stabilization of a Mini Rotorcraft Having Four Rotors, *Control Systems Magazine*, Vol. 25, No. 6, 2005, pp. 45-55.
- [13] Aircraft X650 Quad-rotor, <http://www.infmetry.com/coolstuff/xaircraftx650-quadcopterquadrotor/>
- [14] Koifman, M., Bar-Itzhack, I.Y. "Inertial navigation system aided by aircraft dynamics". *IEEE Trans. Control Syst. Technol.* 1999, 7, pp. 487-493.
- [15] Wang, J.H., Gao, Y. Land, "Vehicle Dynamics-Aided Inertial Navigation". *IEEE Trans. Aerosp. Electron. Syst.* 2010, 46, pp. 1638-1653.

Clustering Multiple Datasets Under Parameter Similarity Constraints

Nikola S. Milosavljević*, Dušan Đ. Okanović**

* Institute for Formal Methods in Computer Science, University of Stuttgart, Germany

** Faculty of Technical Sciences, University of Novi Sad, Serbia

Abstract—Clustering algorithms usually have one or more parameters that control the scale at which the algorithm looks at the data. We study the problem of simultaneously selecting parameter values for multiple datasets (clustering instances), some of which are a priori known to have similar values. We propose two optimization problems related to this task. We show that one of them is NP-hard, and give a polynomial-time algorithm for the other.

Keywords—About four key words or phrases in alphabetical order, separated by commas.

I. INTRODUCTION

Clustering high-dimensional data is a widely studied problem with numerous applications. The goal is to partition a set of high-dimensional points into subsets (*clusters*) so that, generally speaking, the distances within one cluster are smaller than distances between points in different clusters. Of course, depending on the application, one can think of many ways to make this problem precise. For example, clusters can be allowed to overlap or not, the desired number of clusters may be known or unknown, inter- and intra-cluster distances may be defined as maximum, average, root mean square distances, etc.

Common to most clustering algorithms (at least the practical ones) is that they do not attempt to give only one definitive answer, but leave a few parameters that can be fine-tuned when the algorithm is used in a specific application. These parameters control the desired size, separation, or number of clusters. In this paper we restrict our attention to algorithms with *single parameter*, which we call *scale*, because they typically define the scale of the features that we consider relevant for the application, and wish our algorithm to discover.

An example of such an algorithm is one variant of the classical *hierarchical* (or *single linkage*) clustering [5]. One starts with each element as its own cluster, and repeatedly merges clusters whose distance is p times bigger than their diameters (or some function thereof). In this case the parameter is p . Its “optimal” value varies among datasets and applications.

In applications, parameter values are often chosen by trying many values from a certain range, and choosing those that are the most “robust”, in the sense that they can be changed by a large amount without significantly affecting the result. In other words, if we partition the range of parameters into subranges according to the clustering produced, the “correct” clustering corresponds to the largest subrange.

In the hierarchical clustering example, the above idea may be realized as follows. One runs the algorithm until there is only one cluster left (i.e., everything is merged). To each cluster merge event one attaches the corresponding distance-to-diameter ratio. In real-world datasets, the ratio usually experiences a sharp increase between the time when the algorithm forms all

“real” clusters, and the time when it starts merging them. Then ρ is set to lie somewhere in this interval.

In this paper we consider applications where we cluster *multiple datasets*, some of which are *correlated*, that is, known *a priori* to have similar scales. This can happen, for example, when the datasets come from the same experiment, or when nearby “pieces” of the same dataset are examined (see the MAPPER example below). Such correlations provide additional information for scale selection. Now we would like to select scales that are both robust within a single dataset, but also similar among correlated datasets. In other words, when deciding between several clusterings of similar robustness, we should opt for that which respects correlations the most.

II. RELATED WORK

Direct motivation for our work is the MAPPER algorithm for structural analysis of high-dimensional data, proposed by Singh, Mémoli and Carlsson [8, 6]. The input of MAPPER is a high-dimensional dataset, and its output is a simplicial complex which serves as a simplified, topology-preserving representation of the dataset.

We give a brief description of the algorithm (see Figure 1 for an illustration). Define a (scalar or vector) continuous¹ function on the input. Cover the function’s range by topologically simple, partially overlapping patches. Define level-sets as the preimages of these patches. Clustered each *levelset*, using any clustering algorithm. For each cluster create one vertex in the output complex. Create a k -simplex for any $k + 1$ vertices whose corresponding clusters have nonempty intersection (note that such clusters necessarily come from overlapping levelsets).

The algorithm works best if the function is nearly injective, i.e., close to a true *parameterization*. In that case the levelsets are topologically simple, i.e., the clustering step produces one cluster. Since for high-dimensional datasets this is not always easy (indeed, one of the goals of data analysis is to discover such parameterizations), the clustering step is introduced to allow the user to be more “sloppy” in defining the function. With this step, the algorithm works even when levelsets are not simple, but consist of several simple pieces far away from each other; the pieces are detected and separated by the clustering algorithm. The authors suggest a few functions that they expect to work for many datasets: local density, local eccentricity, eigenvectors of the graph Laplacian operator on the data etc.

How does this relate to our work? The datasets that need to be clustered are the levelsets. The levelsets that come from overlapping patches in the parameter space can be expected to have the similar scales, since the function is continuous.

In the original paper about MAPPER [8], Singh, Mémoli, and Carlsson use single-linkage clustering independently for each dataset, i.e., do not exploit spatial coherence of scale at all. In a later version of their paper (only published as poster [7] to our knowledge), they propose a way to take scale coherence into account, but only for scalar functions whose image is a segment on the real line. In this case, the correlated datasets (levelsets) are exactly those that appear consecutively on the real line. In other words, the correlation graph (to be defined precisely below) is a path graph, so their scale selection problem can be solved by computing a

¹ With an appropriate definition of continuous for the finite domain.

shortest path. We suspect that they did not pursue this further because the associated optimization becomes difficult for arbitrary correlation patterns. In fact, in this paper we prove it NP-hard (Section 6).

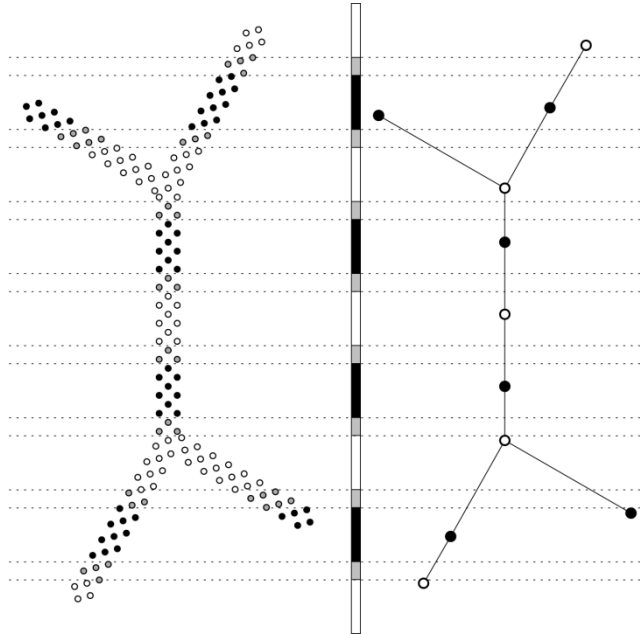


Fig. 1. Illustration of MAPPER. Input dataset is shown on the left. The function is scalar, and its range is a segment on the real line, shown as vertical bar in the center. The output is shown on the right. In this case the output is a graph (simplicial complex of dimension 1). Colors encode covering of the function's range by black and white intervals (center), corresponding levelsets (left) and clusters (right). Gray always represents overlap of black and white.

III. OUR CONTRIBUTION

In this paper we study the scale selection problem for arbitrary correlation patterns. We propose two natural formalizations of the aforementioned scale selection problem. One of them (see PROBLEM B below) is a direct generalization of the one studied in [8]. Unfortunately, we can show that this problem is NP-hard, and hence unlikely to be solvable in polynomial time. Our second formalization (see PROBLEM A below) is more tractable. We give a polynomial time algorithm that solves the problem exactly, and works for *arbitrary* correlation patterns.

The rest of the paper is organized as follows. In Section 4 we define the two optimization problems of interest. In Sections 5 and 6 we present an algorithm for PROBLEM A, and a hardness proof for PROBLEM B, respectively. We conclude the paper in Section 7 by presenting some directions for future work.

IV. PROBLEM DEFINITIONS

In this section we study two combinatorial optimization problems which formalize the aforementioned task of automatic scale selection.

The input is the same for both problems. It consists of:

1. An undirected graph $G = (V, E)$.
2. For each $v \in V$, a nonempty set of real values $S_v = \{s_{v1}, s_{v2}, \dots, s_{v|S_v|}\}$, where $s_i < s_{i+1}$ for all $1 \leq i < |S_v|$.

Each $v \in V$ corresponds to a dataset. $(u, v) \in E$ if the datasets associated with u and v are correlated, i.e., if their scale difference should be low. $[s_{v1}; s_{v|S_v|})$ is the range of scales relevant for the dataset associated with $v \in V$. $[s_{v,i}, s_{v,i+1})$ for $1 \leq i < |S_v|$ are maximal intervals in which clustering of that dataset does not change. Note that if $x \in S_v$, then x belongs to the interval “above” it.

Next we describe the objective functions. For $v \in V$, we denote by $l_v(x)$ the length of the interval in S_v containing value x , that is

$$l_v(x) = \min\{s \in S_v \mid s \leq x\} - \max\{s \in S_v \mid s > x\}.$$

For $u, v \in V$, let $d_G(u, v)$ denote the graph distance (length of a shortest path) in G between u and v . For $p \in S_u, q \in S_v$, we define $d_G(p, q) = d_G(u, v)$.

PROBLEM A We want to select one scale per vertex so as to maximize the *total length of the intervals containing the selected scales*, while respecting an *upper bound on maximum distance between correlated scales*. Formally,

$$\max_{\{x_v \mid v \in V\}} \sum_{v \in V} l_v(x_v) \tag{1}$$

$$\text{s.t. } |x_u - x_v| \leq 1 \quad \forall (u, v) \in E, \quad s_{v1} \leq x_v < s_{v|S_v|} \quad \forall v \in V.$$

Notice that we have assumed an upper bound of 1 on the distance between correlated scales. This is without loss of generality, since we can multiply all values in the instance by the same suitable constant. The last set of constraints ensures that in any feasible solution $l_v(x_v)$ is finite for all $v \in V$.

For $(u, v) \in E$, $1 \leq i < |S_u|$, $1 \leq j < |S_v|$, we denote by o_{uivj} the overlapping length of intervals $[s_{ui}, s_{u,i+1})$ and $[s_{vj}, s_{v,j+1})$, that is

$$o_{uivj} = \max\{\min\{s_{u,i+1}, s_{v,j+1}\} - \max\{s_{ui}, s_{vj}\}, 0\}.$$

PROBLEM B We seek to maximize the total overlapping length of selected intervals for all pairs of correlated datasets. Formally,

$$\max_{\{i_v \mid 1 \leq i_v < |S_v|, v \in V\}} \sum_{(u,v)} o_{u i_u v i_v}$$

Singh, Mémoli and Carlsson [8] considered the very same objective, but they required that G be a line graph.

V. ALGORITHM FOR PROBLEM A

We solve **PROBLEM A** by reducing it to **MINIMUM CUT**, which is well known to be efficiently solvable.

MINIMUM CUT Given a directed graph $G = (V, E)$ with distinguished $s, t \in E$, and non-negative edge capacities $\{c(e) \mid e \in E\}$, find $S \subseteq V$ that such that

$$\sum \{c(e) \mid e = (u, v) \in E, u \in C, v \in C\}$$

i.e., total capacity of the edges from C to $V \setminus C$, is minimized.

The first step is to show that we only need to consider a finite, discrete set of scales.

Lemma 1. *There is an optimal solution $\{x_v \mid v \in V\}$ such that for all $v \in V$ it holds that $|x_v - s| = d_g(u, v)$ for some $u \in V$ and $s \in S_u$.*

Proof. Assume for contradiction that every optimal solution to PROBLEM A violates the aforementioned condition. Fix an arbitrary ordering of vertices and consider the lexicographically smallest optimal solution to PROBLEM A with respect to this ordering. Let v be a vertex that violates the aforementioned property. Let $U = \{u \mid x_u = x_v + d_g(u, v)\}$, i.e., the set of vertices that can be reached from v in G by always increasing the selected scale by 1. Notice that $v \in U$.

We prove that we can decrease $\{x_u \mid u \in U\}$ by a nonzero amount without violating optimality. If this is not the case, then some $u \in U$ either crosses into a different interval (possibly changing the objective function value), or becomes farther than 1 from $x_{u'}$ for some $u' \in U \setminus V$ such that $(u, u') \in E$ (violating feasibility). The first case contradicts the fact that $v \in U$. The second case contradicts $u' \notin U$.

Decreasing $\{x_u \mid u \in U\}$ leads to an optimal solution lexicographically smaller than the one chosen initially, which is impossible.

We define P_v to be the discrete sets from Lemma 1.

$$P_v = \bigcup_{u \in V} \{s + d(u, v), s - d(u, v) \mid s \in S_u\}.$$

We define $P = \coprod_{v \in V} P_v$ (note the disjoint union²).

Now we prune P to make sure that all its elements belong to one of the original scale ranges. That is, we discard all $p \in P_v$ which is not in the interval $[s_{v1}, s_{v|S_v|}]$. We prune P further to ensure that all its elements “connect in all directions”. In other words, we discard all $p \in P_v$ such that $|p - q| > 1$ for all $q \in P_u$, where $(u, v) \in E$. Both types of pruning do not change the optimal solution, because the discarded elements could not have participated in any feasible solution.

If after this pruning step some P_v is empty, then we can immediately declare the PROBLEM A instance infeasible. Otherwise, from now on we can assume the following.

Assumption 1. For all $v \in V$, P_v is nonempty, and all lengths $l_v(p)$ are positive and finite. For any $(u, v) \in E$ and any $p \in P_u$ there exists $q \in P_v$ such that $|p - q| \leq 1$.

Let C be the sum of the longest intervals in for all $v \in V$.

² We keep track of which P_v each element of P came from. In particular, we don't identify duplicates in P .

$$C = \sum_{v \in V} \min_{s \in P_v} l_v(s).$$

Then for any $v \in V$, $p \in P_v$ we define the weight of p to be $w(p) = C + l_v(p)$. For $X \subseteq P$, we define $w(X) = \sum_{p \in X} w(p)$.

Now we claim that can rewrite PROBLEM A as follows.

$$\begin{aligned} & \max_{X \subseteq P} \sum_{p \in X} w(p) \\ & \text{s.t } |p - q| \leq d_G(p, q) \quad \forall p, q \in X \end{aligned} \tag{2}$$

Lemma 2. Any optimal solution to (2) of weight W is a feasible solution to PROBLEM A of length $W - |V|C$.

Proof. Let X be an optimal solution to (2) of weight W . It suffices to show that $|X \cap P_v| = 1$ for all $v \in V$.

It is clear that X cannot have more than one element from some P_v , as this is prevented by the constraints of (2). Therefore, it suffices to prove that $|X| \geq |V|$.

Define $y_v = \min_{p \in P_v} l_v(p)$ for all $v \in V$, and $Y = \{y_v | v \in V\}$. Note that this is well defined because all P_v are nonempty (Assumption 1). Y is feasible for (2). This is because $|y_u - y_v| > 1$ for some $(u, v) \in E$, would imply that either y_u is not connected to any element of P_v , or vice versa, which is impossible by Assumption 1. It follows that $w(Y) > |Y|C = |V|C$.

On the other hand, $w(X) = |X|C + \sum_{p \in X} l(p)$. Since X has at most one element from each P_v , the second term is at most C , so $w(x) \leq (|X| + 1)C$. Since X is optimal for (2), $w(x) \geq w(Y) > |V|C$. This implies $|X| + 1 > |V|$, i.e., $|X| \geq |V|$, as required.

Lemma 3. Any feasible solution to PROBLEM A of length L that satisfies the condition of Lemma 1 is a feasible solution to (2) of weight $L + |V|C$.

Proof. Let X be a feasible solution for PROBLEM A of length L , and suppose X satisfies the condition of Lemma 1. The latter implies that any element of X also exists in P . Let $p, q \in X$, and suppose $p \in P_a$, $q \in P_b$ for some $a, b \in V$. To see that $|p - q| \leq d_G(p, q)$, applying the constraint $|x_u - x_v| \leq 1$ of (1) to every edge (u, v) on some shortest path between a and b . Hence X is feasible for (2).

Lemma 4. (2) has the same optimal solution as (1).

Proof. Let L be the optimal length of (1). By Lemma 1 and Lemma 3, there is a feasible solution to (2) of weight $L + |V|C$. Hence the optimal weight of (2) is at least $L + |V|C$. It cannot be strictly higher than $L + |V|C$, since by Lemma 2, that would imply a solution to (1) of length strictly more than L .

In light of Lemma 4, from now on we refer to (2) as PROBLEM A.

The next step is to map the instance of PROBLEM A to the following instance of MINIMUM CUT (Figure 2). For each $p \in P$, create two nodes p^+ and p^- . Let P^+ (resp. P^-) be the set of all p^+ (resp. p^-), $p \in P$. For any $(p, q) \in P$, connect p^+ and q^- , with infinite capacity, if and only if $p > q + d_G(p, q)$. Create two additional nodes s and t . Connect s to each $p^- \in P^-$ with capacity $w(p)$. Connect each $p^+ \in P^+$ to t , also with capacity $w(p)$. The problem is to find an s - t cut of minimum capacity. Let $W = \sum_{p \in P} w(p)$.

Lemma 5. *If there is a feasible solution with weight l , then there is a cut of capacity at most $W - \lambda$.*

Proof. Let X be an optimal solution. Clearly, $w(X) \geq \lambda$.

Define

$$C = \{s\} \cup \{p^+ \mid \exists x \in X, p \geq x + d_G(x, p)\} \\ \cup \{p^- \mid \forall x \in X, p > x - d_G(x, p)\}.$$

Refer to Figure 3 for a more pictorial representation of this construction.

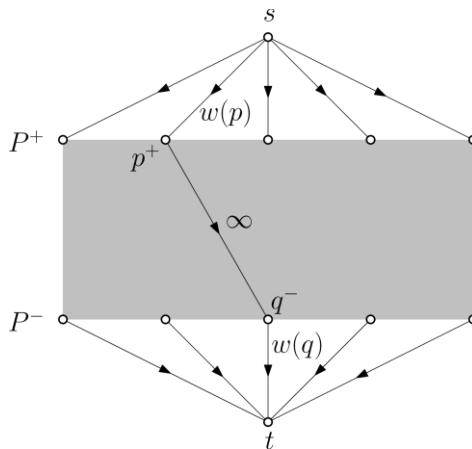


Fig. 2. Solving PROBLEM A via MINIMUM CUT. Edge labels denote capacities.

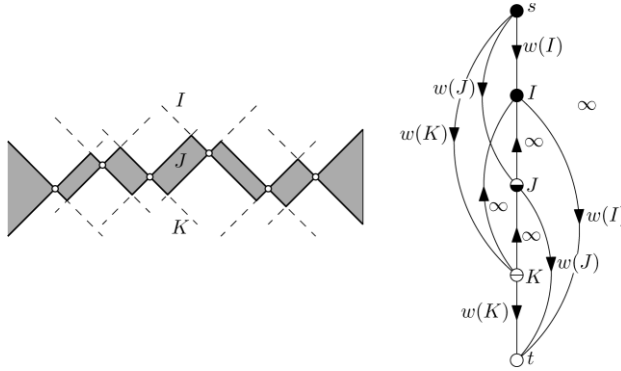


Fig. 3 Constructing a cut from a feasible solution to PROBLEM A. Left: The vertices shown are elements of X . (I, J, K) is a partition of $P^+ \cup P^-$ into “bigger than X ”, “around X ” and “smaller than X ” values. Right: Schematic representation of different parts of the bipartite graph from Figure 2 that appear in the proof. Cut C is represented in black. Upper (resp. lower) “halves” of vertices represent subsets of P^+ (resp. P^-).

Let $p^+ \in C, q^- \in C$. By definition of C , we have

$$\exists x_p \in X, p \geq x_p + d_G(x_p, p) \quad \exists x_q \in X, p \leq x_q + d_G(x_q, q).$$

By feasibility of $X, |x_p - x_q| \leq d_G(x_p, x_q)$. Applying triangle inequality

$$p \geq q + d_G(x_p, p) + d_G(x_q, q) - d_G(x_p, x_q) \geq q - d_G(p, q),$$

hence (p^+, q^-) is not an edge.

It follows that the capacity of C is equal to the capacity of (s, p^+) , for all $p^+ \in C$, and (p^-, t) , for all $p^- \in C$. The latter is equal to

$$\sum \{w(p) \mid \forall x \in X, p < x + d_G(x, p)\} + \sum \{w(p) \mid \forall x \in X, p > x - d_G(x, p)\}. \quad (3)$$

Some $p \in X$ cannot appear in either of the two sums, because the respective condition fails for $x = p$. Now suppose some $p \notin X$ appears in both the sums. Then

$$\forall x \in X, |p - x| < d_G(x, p),$$

so $X' = X \cup \{p\}$ is feasible. But then $w(X') > w(X)$, which contradicts the choice of X . We conclude that $p \notin X$ can appear at most once in (3).

Since any $p \notin X$ is counted at most once, and any $p \in X$ is not counted at all, (3) evaluates to at most $W - w(X) \leq W - \lambda$.

Lemma 6. *If there is a cut of capacity γ , then there is a feasible solution with weight at least $W - \gamma$.*

Proof. Let $C \ni s$ be a cut of minimum capacity. Clearly, the capacity of C is at most γ . Define

$$X = \{p | p^+ \in C, p^- \notin C\}. \quad (4)$$

For each $p^+ \in C, p^- \notin C$, we have that (p^+, q^-) is not an edge; otherwise $C \cup \{q^-\}$ would be a cut of smaller capacity than C , contradicting the choice of C . The capacity of C is therefore equal to the capacity of (s, p^+) , for all $p^+ \in C$, and (p^-, t) , for all $p^- \notin C$.

We prove that X is feasible. Consider any $p, q \in X$. By definition of X , we have $p^+, q^+ \in C, p^-, q^- \notin C$. Since p^+ is not adjacent to q^- , it holds $q \leq p + d_G(p, q)$. Since q^+ is not adjacent to p^- , it holds $p \leq q + d_G(p, q)$. Hence $|p - q| \leq d_G(p, q)$, as required.

We lower bound the weight of X .

$$\begin{aligned} w(X) &= \sum \{w(p) | p^+ \in C, p^- \notin C\} = \sum \{w(p) | p^+ \in C\} - \sum \{l(p) | p^+ \in C, p^- \in C\} \\ &\geq \sum \{w(p) | p^+ \in C\} - \sum \{l(p) | p^- \in C\} \\ &= W - \left[\sum \{w(p) | p^+ \notin C\} + \sum \{l(p) | p^- \in C\} \right]. \end{aligned}$$

The term in brackets is equal to the capacity of C , which is at most γ by assumption.

Theorem 1. PROBLEM A can be solved in time polynomial in the size of the input.

Proof. The algorithm is

1. Compute the discrete set of scales P .
2. Compute an instance of MINIMUM CUT.
3. Solve it to get a minimum cut C .
4. Return solution X computed from C according to (4).

Correctness of this algorithm directly follows from Lemma 5 and Lemma 6. We analyze the running time. Let $|V| = n, |E| = m, \sum_{v \in V} |S_v| = s$. Initial input size is $O(m + s)$. Step 1 can be implemented in $O(ms)$ time (one breadth-first search from each node), producing a set of size $O(ns)$. As a byproduct, step 1 can also compute and store $d_G(u, v)$ for all $u, v \in V$. Step 2 outputs a bipartite graph with $O(ns)$ vertices and $O((ns)^2)$ edges in $O((ns)^2)$; implementation of this is straightforward. Step 4 trivially takes $O(ns)$ time. Hence steps 1, 2 and 4 together run in $O(ms + (ns)^2) = O((ns)^2)$ time. Step 3 is a minimum cut computation on a graph with $O(ns)$ vertices and $O((ns)^2)$ edges. It can be implemented in $O((ns)^3)$ time using [4]³. Hence the running time for the whole algorithm is $O((ns)^3)$.

VI. HARDNESS OF PROBLEM B

In this section we show that PROBLEM B is at least as hard as MAXIMUM INDEPENDENT SET, even when all intervals are of fixed length l and if their endpoints are multiples of l .

³The algorithm of Goldberg and Rao [3] gives a theoretical running time of $O((ns)^{2/2})$ for step 3, but it is more complicated to implement, and probably slower in practice.

MAXIMUM INDEPENDENT SET Given an undirected graph, find a subset of maximum cardinality whose vertices are pairwise non-adjacent.

Let $H = (V_H, E_H)$ be the graph from the instance of MAXIMUM INDEPENDENT SET. We construct an instance of PROBLEM B, that is, a graph $G = (V_G, E_G)$ and sets $S_v, v \in V_G$. Please refer to Figure 4 for illustration. Let $V_H = \{u_1, u_2, \dots, u_{|V_H|}\}$. For each u_i , add to V_G a vertex v_i with $S_{v_i} = \{(2i - 2)l, (2i - 1)l, 2il\}$, and a vertex z_i with $S_{z_i} = \{(2i - 1)l, 2il\}$. Add (v_i, z_i) to E_G for all i . For edge $(u_i, u_j) \in E_H$, add a vertex $v_{ij} \in V$ with $S_{v_{ij}} = \{0, l, 2l, \dots, 2|V_H|l\}$. Add (v_{ij}, v_i) and (v_{ij}, v_j) to E_G . Let $K = 2|E_H| + V_H$. Add v_{ijk} to V_G , where $1 \leq k \leq 2K$, and $S_{v_{ijk}} = \{(2i - 2)l, (2i - 1)l\}$ for $k \leq K$, and $S_{v_{ijk}} = \{(2j - 2)l, (2j - 1)l\}$ otherwise. Add (v_{ijk}, v_{ij}) to E_G for all k .

Theorem 2. PROBLEM B is NP-hard.

Proof. It follows directly from Lemma 7 and Lemma 8 below that H has an independent set of size α if and only if there is a feasible solution to the associated PROBLEM B instance G with objective value of $[|E_H|(K + 1) + \alpha]l$. Since G can be computed from H in polynomial time, and MAXIMUM INDEPENDENT SET is NP-hard on general graphs, the claim follows.

Lemma 7. If H has an independent set of size a , then there is a feasible solution to PROBLEM B instance G with objective value $[|E_H|(K + 1) + \alpha]l$.

Proof. For v_i select interval $[(2i - 1)l, 2il]$ if u_i is in the independent set, and interval $[(2i - 2)l, (2i - 1)l]$ otherwise. For v_{ij} , select interval $[(2i - 2)l, (2i - 1)l]$ if u_i is not in the independent set, and interval $[(2j - 2)l, (2j - 1)l]$ otherwise. For v_{ijk} and z_i select the only available interval.

Along the edges of the type (v_{ij}, v_{ijk}) , for fixed i, j , there are exactly K overlaps. Along the two remaining edges (v_{ij}, v_i) and (v_{ij}, v_j) there is exactly one overlap. Along the edges of the type (v_i, z_i) , summed over all i , there are exactly a overlaps. The claim follows by adding the above counts.

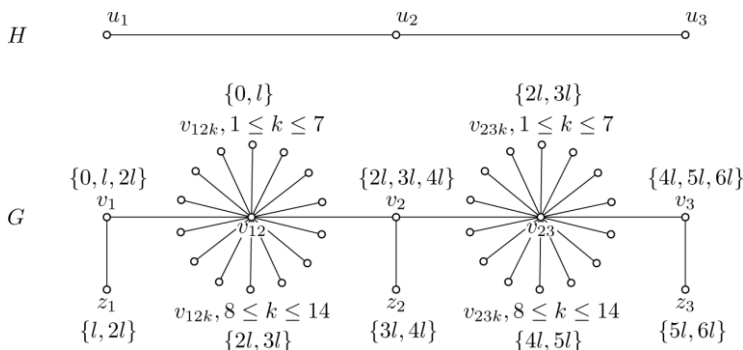


Fig. 4. Illustration of mapping from H to G when H is the path graph of length 2. The set next to each node label

v is S_v .

Lemma 8. *If there is a feasible solution to PROBLEM B instance G with objective value at least $[|E_H|(K + 1) + \alpha]l$, then there is an independent set of size at least α .*

Proof. Suppose that the premise is true. Then there is an optimal solution with objective value at least $[|E_H|(K + 1) + \alpha]l$. Clearly, it has at least $|E_H|(K + 1) + \alpha$ overlaps.

Notice that edges incident to a node of type v_{ij} contributes at most $K + 2$ overlaps. However, if some v_{ij} selects neither $[(2i - 2)l, (2i - 1)l]$ nor $[(2j - 2)l, (2j - 1)l]$, then all edges incident to it contribute at most two overlaps. Consequently, the number of overlaps for the whole graph is at most $(|E_H| - 1)(K + 2) + 2 + |V_H|$.

On the other hand, if all v_{ij} select one of those two special intervals, then the number of overlaps for the whole graph is at least $|E_H|K$. Since $K > 2|E_H| + |V_H|$, the latter strictly exceeds $(|E_H| - 1)(K + 2) + 2 + |V_H|$. Hence we have proved that in any optimal solution all v_{ij} must select one of the special intervals. Now we have that the edges of type (v_{ij}, v_{ijk}) contribute exactly $|E_H|K$ overlaps.

Next we consider the edges of the type (v_{ij}, v_i) and (v_{ij}, v_j) . They contribute at most $|E_H|$ overlaps, because each pair of edges incident to the same v_{ij} contributes at most one. Now we modify the solution without affecting feasibility or changing the objective. The goal is to have each pair of edges $(v_{ij}, v_i), (v_{ij}, v_j)$ contribute exactly one overlap. This can be accomplished as follows. Pick any pair $(v_{ij}, v_i), (v_{ij}, v_j)$ that violates this property, and change the selection for either v_i or v_j , to match the current selection of v_{ij} . This clearly does not affect feasibility or change the objective (at most one overlap is lost, the one with z_i or z_j , and at least one is gained, the one with v_{ij}). It is also obvious that it decreases the number of edge pairs that violate the property by one. So if we repeat this modification enough times, each edge pair $(v_{ij}, v_i), (v_{ij}, v_j)$ will eventually contribute one overlap.

That leaves at least α overlaps for the edges of the type (v_i, z_i) . We prove that exactly those overlaps induce an independent set in H . In other words, the set $X = \{u_i | (v_i, z_i) \text{ contributes an overlap}\} \subseteq V_H$ is independent. This is not hard to see; if $u_i, u_j \in X$ were adjacent in H , then the corresponding edges $(v_i, v_{ij}), (v_{ij}, v_j)$ in G would not contribute any overlaps.

VII. FUTURE WORK

The idea of examining a range of parameters and accepting the one that is most stable can be applied to discovering not only connected components (clusters) in a dataset, which was the main topic of this paper, but also loops, “voids” and higher-dimensional topological features. This has been formally studied as *persistent homology* [1, 2]. In this framework, each element

of S_v , would have additional attributes, such as a flag that says whether a topological feature was created or destroyed (both are possible) for the respective value of the parameter, as well as the *dimension* of the feature. One direction for future work is to extend our algorithm so that it takes these extra attributes into account.

Approximation algorithms for Problems A and B are an interesting topics for future work. Unfortunately, neither of the two reductions presented in this paper is approximation preserving, because the additive difference between the two objectives can far exceed one of the objectives. As a consequence, we cannot gain by using faster and simpler-to-implement algorithms for computing approximately minimum cuts. Likewise, the reduction in Section 6 does not allow us to exploit strong inapproximability of MAXIMUM INDEPENDENT SET to rule out efficient approximation algorithms for PROBLEM B.

REFERENCES

- [1] H. Edelsbrunner and J. Harer. Persistent homology - a survey. *Surveys on Discrete and Computational Geometry. Twenty Years*. 2008, pp. 257-282.
- [2] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, vol. 28, 2002, pp. 511–533.
- [3] A. V. Goldberg and S. Rao. Beyond the flow decomposition barrier. *J. ACM*, vol. 45, September 1998, pp. 783–797.
- [4] A. V. Goldberg and R. E. Tarjan. “A new approach to the maximum flow problem.” In *Proceedings of the eighteenth annual ACM Symposium on Theory of Computing, STOC '86*, New York, NY, USA, 1986, pp. 136–146.
- [5] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, vol. 2, 1967, pp. 241–254.
- [6] G. Singh. Algorithms for topological analysis of data. PhD thesis, Stanford University, 2008.
- [7] G. Singh, F. Mémoli, and G. Carlsson. “Mapper: A topological method for analysis of high dimensional data sets.” poster in *Topology Learning*, a NIPS 2007 workshop, 2007.
- [8] G. Singh, F. Mémoli, and G. Carlsson. “Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.” In *SPBG07-proc*, 2007, pp. 91–100.

Realistic Terrain Aware Mobility Model

Maja Dineska, Sonja Filiposka*

* Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Macedonia

Abstract— Among other simulation parameters, topology and mobility model are key factors for precise evaluation of ad hoc networking protocol characteristics. Because the movement of the nodes directly impacts protocol performance, it is essential to use realistic movement model that will provide improved simulation results. The majority of proposed mobility models in the current research literature do not provide realistic movement scenarios in a terrain modeled environment. They are often limited to random walk mobility models without considering real-world terrain. The work presented in this paper introduces R3D, an independent tool for generating mobility scenarios that follow a newly proposed 3D mobility model that includes nodes movements through real, irregular 3D terrain.

Keywords—3D, mobility model, terrain.

I. INTRODUCTION

In the next generation of wireless communication systems, there will be a need for the rapid deployment of independent mobile users. Significant application examples include establishing efficient, dynamic communication for emergency operations, disaster recovery or military operations. Such network scenarios do not rely on existing communication infrastructure and can be conceived as applications of Mobile Ad Hoc NETWORK [1]. A MANET is composed of mobile devices capable of wireless communication, such as user-carried PDA devices and notebooks. Unlike wired networks which rely on routers or managed (infrastructure) wireless networks that rely on access points, wireless ad hoc networks perform routing by forwarding data between nodes. All mobile nodes act as mobile routers. Ad hoc networking protocols depend heavily on the routing mechanism and the movement of the nodes.

Network simulations are quite often the only research tool for understanding the operation of ad hoc networking protocols. Simulations provide feasible mean to compare different protocols and analyze their performance. Indeed, network simulations environments such as NS-2 [2], Qualnet [3] or Opnet [4] are the most commonly used tools for performance evaluation. There are a number of important simulations parameters, but topology and mobility models are two of the key factors for obtaining acceptable realistic results [5].

One important feature of MANETs is the dynamic behavior caused by node mobility. Thus, key challenge in the evaluation of such protocols is to conduct the performance analysis with realistic mobility models that accurately reflect the mobile users' movement. A realistic mobility model should include avoiding and getting around real-world obstacles that will provide conducting much improved analysis over real life scenarios.

Mobility models describe the movement pattern of mobile nodes and provide a definition for their location and velocity that might change over time. Once the nodes are initially placed, the mobility model is the one that defines movements over the simulation area. Many mobility models for the generation of synthetic traces have been presented (a survey is provided by

Camp, Boleng and Davies in [6]). The first mobility model ever proposed was the Random Waypoint [7] movement model. Although it is the first it is also still the most widely used model. However the simulated behavior using this model does not resemble the natural movement of the nodes and point to several weaknesses [8]. Therefore, in attempt to improve movement patterns and to increase the realistic features, various researchers proposed large set of mobility models with different characteristics [9] [10].

The firstly proposed mobility models are the entity mobility models that are concerned with the individual node's movement. This individual movement is entirely independent of the movements of other nodes and the environment, although its changes in direction and speed in time interval $(t+1)$ may depend on their values in the previous time interval t [10]. Models that demonstrate this feature are said to be models with temporal dependency [11]. A step forward towards realistic models is the ability of nodes to cluster and therefore the models are referred to as Group Mobility Models [12]. This means that node's movement may be influenced by neighboring nodes (e.g. ad hoc meetings, instant information sharing, etc.). Topographical models [13] [14] on the other hand integrate the environment in the simulation area following the necessity that the node's movement must be restricted by topographical characteristics.

It is fairly straightforward to conclude that all entity movement models generate behavior that is most inhuman-like. Having in mind that the most likely deployed scenarios for mobile ad hoc networks are found in outdoor scenarios for rescue missions, exploration and similar, one must emphasize the need for realistic mobility model that will integrate the environment which on the other hand can significantly restrict the movement of the nodes as well as the propagation of the wireless signals.

Based on this analysis, in this paper we propose a standalone tool that implements a newly defined realistic 3D mobility model that aims to realistically model the node mobility throughout a real three-dimensional terrain with realistic obstacles. The tool generates an output that defines the movement of the nodes during the simulation time in a format that can further be used as a mobility scenario script for the NS-2 simulator.

The remainder of the paper is organized as follows. Section 2 details related research in the area of mobility models. Section 3 describes the definition of the terrain over layer while the design and implementation of the model is presented in Section 4. Analysis of the properties of the model are provided in Section 5. Finally, Section 6 concludes the paper, outlining the future research direction.

II. RELATED WORK

There is a lot of literature that deals with the properties and descriptions of the proposed mobility models and their movement patterns. Categorization of mobility models can be found in [11], while a survey and a general comparison is provided in [6]. In this section we provide a brief description and an overview of the benefits and deficiencies obtained with the use of different mobility models and argue what is the missing puzzle in order to achieve more realistic behavior.

Random based individual movement models are still the most widely used and not because of their correct results, but primarily because of their simplicity. The Random Waypoint model [7] is used on a large scale when simulating protocols designed for mobile ad hoc networks. Because of its simplicity it is available with a large set of simulators. Mobile nodes move

randomly in a two-dimensional system area without any restrictions in terms of the environment. In addition, parameters such as destination, speed and direction are all picked randomly. In brief, the order of the actions is as follows: each node picks a random destination uniformly and travels with speed v whose value is uniformly chosen in the interval $(0, V_{max}]$. When a node reaches its destination point, it takes some pause time, after which it chooses a new destination and a new speed and resumes movement. The advantage of the model is its simplicity in implementation and performance evaluation and therefore is the most commonly used for simulation purposes.

On the other hand several studies [15] of this model have revealed unreliable results and many deficiencies. This model is expected to maintain the average speed as the simulation progresses, however in [8] it is shown that this model fails in providing steady state in a sense that the average speed of a mobile nodes constantly decreases with time. This is due to the fact that more and more nodes are “stuck” traveling to their destinations at low speeds. It was also shown that nodes distribution is higher in the center of the simulation area compared to the boundaries. Nodes traveling towards their destinations take sharp changes in directions and velocity [16] which on the other hand are chosen without consideration of their previous values. Almost all weaknesses discussed above are true for all random mobility models. The Random Walk Mobility Model [15] can be considered as Random Waypoint with pause time between two movements equal to zero. Random Direction [9] is proposed to overcome the unexpected issues that produce Random Waypoint regarding density waves.

While each of these models generates random mobility and can be used for the simulation study of ad hoc networking protocols, none of these models attempts to model the behavior of nodes in a realistic environment. All of the models assume open, unobstructed areas in which nodes are free to move according to the constraints of the mobility model. In real-world scenarios it is quite uncommon that people are located on flat terrains with no obstructions at all. In order to understand how a protocol will behave in an obstructed area, it is necessary to create mobility models that precisely model the environment.

There are a few mobility models that include the profile of the terrain when constructing the mobility scenarios, but these are often constrained by moving only on horizontal and vertical streets that represent the unobstructed area [14] or on the lanes of the freeway. Although these models incorporate sort of real terrains, they cannot be considered as realistic, because of their simplicity correlated with spatial dependency and inability to provide more complex realistic behavior. The Obstacle Mobility Model considers [13] the real-world urban topography, but only in a two dimensional system. The profile of the terrain used in this model includes obstacles like buildings and other objects that are target destination for the nodes. Several models have been developed based on this Obstacle Mobility Model. However one must bear in mind that while these models incorporate the environment in the mobility scenarios they are all models that try to model the urban node behavior.

On the other hand, because ad hoc network deployment scenarios are often placed in outdoor non urban environments, the R3D mobility model proposed and presented in this paper provides a mechanism for modeling movement in real world outdoor non urban scenarios that is based on irregular three-dimensional terrains defined using a standard terrain description according to the Geographical Information Systems.

III. TERRAIN OVER LAYER FOR MOBILITY

In order to create a 3D terrain aware mobility model the first step is to be able to read in the terrain configuration from a standardized GIS terrain format. Regarding the mobility modeling of outdoor non-urban scenarios, the existing approaches are neither suitable nor complete. In order to achieve high degree of realism geographic restrictions must be considered. Defining geographic restrictions means implementing topology sub-model that influence node movements regarding the terrain.

The terrain definitions that can be used in our model are based on the standard formats for terrain description: Digital Elevation Model (DEM) [17] or Triangular Irregular Networks (TIN) [18]. The R3D tool accepts the TIN terrain description as an input argument as defined through the use of the Virtual Reality Language (VRL) [19]. However since the access to a freely available DEM definition of a real world terrain is much easier, one can use these DEM terrain format with an extra step of converting the DEM to TIN (without any information loss) using some freely available software such as Landserf [20].

The TIN vector model represents a surface as a set of contiguous, non-overlapping triangles associated with 3-dimensional data (x, y, and z) and topography. Within each triangle the surface is represented by a plane. The points that define all the triangles and planes that describe the terrain are the data that the VRL file contains. For a comparison the DEM format is a raster based format that holds the terrain information as a matrix of terrain heights. An advantage of using a TIN over a raster DEM in mapping and analysis is that the points of a TIN are distributed variably based on an algorithm that determines which points are most necessary in order to provide an accurate representation of the terrain. Data input is therefore flexible and fewer points need to be stored than in a raster DEM, with regularly distributed points.

The R3D mobility model is based on the random waypoint mobility model while making the nodes aware of the terrain they are moving on. This terrain awareness is primarily done using a second layer over the terrain that defines the allowed approachable movable areas for the nodes and denies node movement in the forbidden non-approachable areas. This second layer over the terrain is defined in 2D and can be obtained automatically based on the steepness and roughness of the terrain or can be defined by the mobility scenario modeler. The main idea of the model is to transform the simulation area in such a way that will not allow the nodes to climb too steep paths (like canyons or steep rocks), or will not allow the nodes to enter natural obstacles (like rivers, lakes, creeks and alike).

The non-approachable areas for the nodes that are automatically recognized via their high slopes are obtained using the coordinates of the points that define the terrain triangles from the TIN file. Using this coordinates it is fairly straightforward to form a connected flat surface plane that represents the triangle extending in three directions and to calculate its normal vectors a, b and c. After determining the values of these vectors, the slope of the plane can be calculated. After each slope is defined, the input terrain is overlaid with a 2D layer that marks each part of the terrain as approachable or not approachable according to the slope of the terrain and the outside input from the user who defines what is considered as an acceptable slope. The user is allowed to define different slope for each scenario, depending on what are the observed properties and the definition of the scenario by itself. Note that the slope is presented in percents and not in degrees.

The detailed terrain modeling process is as follows. For each cell in the two-dimensional

layer the central cell point equally located from both ends is selected and then the slope of the triangle containing this central point is observed in order to compare it with the acceptable slope range. If the value of the slope is greater than the one defined by the user as maximum acceptable, then the cell is modeled as unapproachable and vice versa. For an example, let's assume that the user defines max acceptable slope of 30%. It means that all the cells for which their central point belongs to triangles with a slope lower than 30% will be marked as approachable and vice versa. The result of this process is a two-dimensional layer representing the terrain that is modeled depending on the three-dimensional description and the user input for the slope.

The described modeling process introduces the problem of too many small approachable areas surrounded by unapproachable areas which allows for the nodes to be stuck in a kind of isolated small islands of non-steep terrain. It is unlikely that one can carry out an efficient analysis and extract proper results, because of the constrained node movements with similar characteristics. The number of moves and pauses performed by the trapped nodes are significantly bigger just as the number of failures realized in order to find a valid destination.

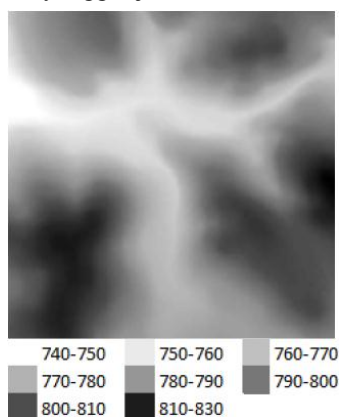


Fig 1. Example terrain and directions for interpreting (height is given in meters)



Figure 2. 2D over layer with approachable and non approachable areas for the example terrain (white approachable; black non approachable areas)

Since it is not a realistic node behavior for the nodes to come to be and remain captured in these areas, this problem was solved in the following way: before the simulation starts, the user is prompted to input the maximum size of the enclosed areas (islands) that will be set as unapproachable. After the initial definition of the over layer with the approachable and non approachable parts of the terrain, the algorithm then includes a calculation that seeks out all enclosed areas and if their surface is smaller than the one defined by the user, it models the whole island as a non approachable terrain. In this manner, the end user can be certain that the nodes can perform movements that are distributed across the whole simulation area, or can be confined on large isolated planes (i.e. mountain plateau).

For a visual example of the construction of the terrain over layer consider the following: the terrain presented in Fig. 1 is a 1000 x 1000 m terrain with a minimum of 730 and maximum of 830 meters height. Using our R3D mobility model we obtain the terrain over layer as it is presented in Fig. 2. The obtained terrain over layer is defined with maximum acceptable slope

of 25% while the minimum acceptable size of the isolated closed areas of 400 square meters.

The inability of the generator to conclude what is the ideal size for manipulation may be conceived as deficiency. The user must perform detail observation of the terrain and maybe use the generator for a couple of times in order to come to the conclusion of the ideal size. Please note that such algorithm is difficult to provide since different scenarios define different parameters and may use many kinds of terrains with varying size. Thus, this feature remains to be improved in the future versions of the mobility model.

The terrain over layer given in Fig. 2 can further be changed by additional markings of non approachable areas by the user. In this way the user can input other non approachable areas like water bodies and alike. Another important remark regarding the R3D mobility model is towards the variety of applications. Please note that the mobility model can actually also be used in a 2D environment only. In this case the user alone defines the non-approachable areas and is not concerned with the accessibility of the steep terrain because it is simply not present in this type of scenario.

IV. DESIGN AND IMPLEMENTATION OF THE MOBILITY MODEL

The R3D mobility model and tool proposed in this paper generates realistic movement pattern for non-urban scenario where irregular terrain is considered. Therefore nodes must follow a proper mobility algorithm avoiding obstacles and only progress towards approachable points.

Based on the over layer of approachable and non approachable areas, the basis of the R3D mobility model is on top of the Random Waypoint mobility model in combination with the A* algorithm [21] used for path finding. This algorithm allows the node to find its way around the non approachable areas towards its goal.

Before the simulation starts, the user is allowed to set a few parameters that will define the scenario. Parameters enabled for user definition are setting the minimum and maximum values for speed and pause (also the user can decide upon enabling pauses for the simulation). The user defines the number of nodes that will be distributed over the simulation area and sets the maximum simulation time. As mentioned in section 3, the size of the maximum acceptable slope and the size of the enclosed areas can be also defined by the user before the simulation starts.

The nodes are initially randomly distributed in the approachable areas. Their potential initial locations are only the cells that are assigned as approachable with the process of terrain modeling. If pauses are enabled by the user, than each node may choose whether to move or pause as its first action. If the pauses are not enabled it is certain that all nodes will perform movement as the simulation begins. If the node will perform movement, it randomly selects one location (cell) in the simulation field as its destination. Then the nodes moves with a randomly chosen speed within a given $[Vmin, Vmax]$ evenly distributed interval towards the randomly chosen goal that has to be positioned in an approachable area while there has to be a path from the node present position to the approachable destination. Valid destination selections are only the cells that are assigned as approachable and there exists a valid path obtained using A* avoiding all the non approachable areas.

If the node's first action is to pause it will only choose a value from the interval $[Pausemin,$

Pausemax] that presents the amount of time that the node will pause after the simulation begins. After the pause time ends nodes are not allowed to perform another pause and are forced to move towards another randomly chosen valid destination in the simulation area with a randomly chosen speed. After each movement the node chooses whether to pause or continue moving toward another reachable randomly chosen goal with randomly chosen speed. The speed and destination goal of a node are chosen independently of other nodes. If the pause time is equal to zero or pauses are not enabled at all, this leads to continuous mobility. The whole process of pausing and moving for all nodes is repeated over and over again until the simulation time reaches the maximum simulation time parameter defined by the user before its start.

As previously commented, the presented mobility model design reminds of the random Waypoint model if the terrain is completely flat. If this is the case, then the tool will act as two-dimensional mobility generator and the movement is not constrained by any obstruction. This leads to the fact that the model may be also used in scenarios that doesn't consider the terrain at all. Regardless of the terrain oblique, we remind that the user is also allowed to define its own obstructions in the simulation field. These obstructions are types of obstructions that somehow are not incorporated with the terrain description. These unapproachable zones will be treated in the same manner as all other cells assigned using the terrain modeling process.

The R3D generator also includes a mechanism for collision avoidance. Nodes traveling towards their destinations may be determined to pass the same cell at the same time and to perform collision. The node collision avoidance works as it is given in [22], but include collision detection before the move even starts. After the node chooses its next destination, the pass time interval is determined for every cell on the path. Then, the node checks if there is another node that will pass any of its cells on the path to the destination in the same time. If this is true, the cells where the collision occurs are assigned as unapproachable for the given time moment and a new path is found. The process is repeated until a path with no collisions at all is found.

V. MOBILITY MODEL CHARACTERISTICS EVALUATION

The primary objective of the simulations given in this paper is to understand the impact of the terrains in a simulation environment. We tested the proposed mobility model using several runs generating different scenarios and making a comparison among the evaluated performances so conclusions could be drawn. Results of multiple scenarios are presented in this section, each one characterized by different parameters, in order to cover all the aspects.

The first group of results is obtained utilizing the square 1000 x 1000 m terrain given in Fig. 1 and appropriately in Fig. 2. The aim of these simulations is to present the impact of the changes in the value of the acceptable terrain slope parameter on the generator performances. We compared the same scenario, first with maximum slope of 30% and second with maximum acceptable slope of 35%. As expected, as the maximum acceptable slope rises, the percent of approachable cells follows. Indeed, 72,3% of the cells in the first run are assigned as approachable despite 87,7% approachable cells after the second run. Our analysis shows that the number of failures to find an approachable destination, and a correct path to it, is directly dependent on the terrain accessibility. As the terrain is more accessible, the less is the number of failures.

The number of collisions avoided is also dependent on the terrain accessibility and also the

number of nodes in the simulation area. If the simulation area is less obstructed, more of the area is approachable and nodes prefer to move across different paths to reach the destinations, so the chance to collide is reduced.

Another parameter that can be indirectly controlled by the user refers to the number of triangles that are used to describe the terrain. The more triangles are involved in the terrain description, the more precise the evaluation gets since the generator works with a more detailed terrain. Through several scenario runs where the same terrain is defined by a different number of triangles we concluded that as the number of triangles decreases the terrain accessibility increases. This is due to the fact that if the terrain description is less precise (in term of number of triangles used to describe the terrain) it means that the average slope of few triangles now merged as one is less than the slope of the previous triangles having the steepest plane and the chance to belong to the defined interval by the user are greater.

For an example, an evaluation was made using another square terrain sized as the previous, with similar characteristics and the following results are obtained: If the terrain is described using 12019 triangles the terrain accessibility is exact 70%, while when the terrain is described using 9843 triangles we get accessibility of 70.5%, while 6862 triangles description that lead to a accessibility of 71.5%. As the number of triangles increases, the number of approachable cells is decreasing leading to more failures in order to find correct destination and more collisions avoided. However, one can observe that the change in accessibility is not very prominent.

It is very important to stress that our analysis has shown that none of the parameters used to define the terrain impact the speed of the nodes. Once the nodes are distributed over the simulation area, each node selects its speed from the given range defined by the user. The speed and the duration of pauses are distributed randomly and uniformly from the given range and their values do not depend on other parameters. As a concluded result from all observations and simulations analysis we may say that the average speed stabilizes a bit lower than the middle of the range defined by the user $(V_{min} + V_{max})/2$. This kind of behavior of the mobility model is expected and encouraged by the existing mobility model analysis [7].

Yet, the number of pauses and the number of movements certainly depend on the duration of the simulation, but also from nodes initial position. If the node is initially located in some enclosed area (island) or maze that is not connected to the rest of the simulation field, then the number of performed movements and pauses will be much greater just as the number of failures to find a valid destination. The bigger the terrain is, the duration of simulation run should be longer in order to have a more even distribution of the visited approachable cells in the simulation field.

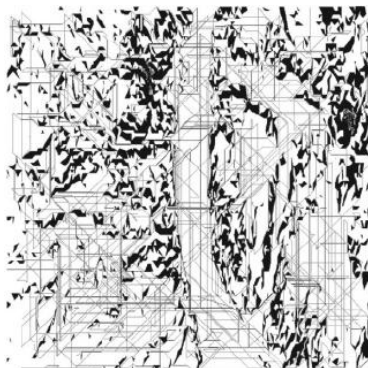


Fig. 3. Movement distribution

Previous simulations using the smaller field of 1000 x 1000 meters were run for 1000 or 1500 seconds. However, if the simulation is run for 1000 seconds incorporating a terrain with size 7000 x 3000 meters, the results show that during this time some of the nodes did not manage to finish even the first movement and some of them barely perform maximum of two movements. The 2D over layer of this terrain is presented in Fig. 3 where are shown all the movements throughout the entire simulation area performed by 50 nodes for 5000 seconds. To conclude, if the tool is about to be used for generating scenario over a large terrain it is a must to perform the simulation several times in order to analyze obtained results. Towards it, we consider that correct results are obtained when the simulation is run long enough to provide velocity stabilization a bit lower than the middle point in the defined range, also for the average pause time (if are enabled). In this way the nodes will perform at least a number of moves and pauses so that correct average numbers may be extracted.

VI. CONCLUSION

Researches indicate that the behavior of simulated routing protocols varies widely depending on the mobility model. The definition of realistic mobility model is crucial since the simulations should provide reliable results for ad hoc networking protocols performance.

It is shown that random mobility models do not provide such results, since the topology map doesn't consider environment at all. Some augmented models as Mobility Model founded on social network theory or 3D signal obstruction model are step forward towards realistic, but still cannot be taken into consideration since they do not implement geographic restrictions and spatial dependency respectively.

This paper proposes a new mobility model that enables inclusion of 3D irregular terrain and generates movement pattern that is in accordance with the terrain profile. The tool may also be used to generate scenarios that incorporate 2D flat terrains, where the area that is approachable is defined by the user only and nodes movement is not also restricted by the terrain profile. In such case, the proposed model will behave just as Random Waypoint model and the model for radio propagation will only consider the distance between nodes.

Regardless of the terrain existence, the user is allowed to define its own unapproachable zones that are not included into the terrain description and to further restrict movements of the nodes.

The R3D mobility model includes a terrain modeling mechanism plus an algorithm for finding the shortest path from the source to destination while providing a mechanism for avoiding all obstructed cells, a mechanism for avoiding collisions and manipulation with small enough closed areas. The results from our simulation analysis show that the model does not have a problem with a continuous decreasing velocity while it is independent on the terrain parameters.

Our future work will be focused on improving the model with a better algorithm for seeking out and defining enclosed unapproachable areas as well as more transparent definition of additional natural or manmade obstacles using a GIS defined over layer of the terrain. Another addition will be incorporating possibilities for social group node behavior in our model.

REFERENCES

- [1] C.Siva Ram Marhy, B. S. Manoj. Ad hoc wireless networks: architecture and protocols. New Jersey:Prentice Hall, 2004
- [2] The Network simulator ns-2. <http://isi.edu/nsnam/ns/>
- [3] Qualnet, www.qualnet.com
- [4] Opnet Modeler. www.opnet.com
- [5] C.P.Agrawal, O.P. Vyas, M.K. Tiwari. "Evaluation of Varying Mobility Models & Network Loads on DSDV Protocol of MANETs", *International Journal on Computer Science And Engineering* Vol.1(2), 2009,pp. 40-46.
- [6] T. Camp, J. Boleng, and V. Davis, "A survey of Mobility Models for Ad hoc Network Research," *Wireless Communication & Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications*, vol.2, no.5, pp. 483-502, 2002.
- [7] D. Johnson and D. Maltz. "Dynamic source routing in ad hoc wireless networks." in *Mobile computing*, T. Imelinsky and H. Korth, Eds., Kluwer Academic Publishers, 1996, pp. 153-181.
- [8] J. Yoon, M. Liu, and B. Noble, "Random Waypoint considered harmful", *Proceedings of IEEE Information Communications Conference (INFOCOM 2003)*, vol 2. Pp. 1312-1321, 2003, CA, USA.
- [9] E.Royer, P.M.Melliar-Smith and L.Moser, "An analysis of the Optimum Node Density for Ad Hoc Mobile Networks", *Proceedings of IEEE International Conference on Communications (ICC)*, 2001.
- [10] B.Liang, and Z.J.Haas, "Predictive Distance-based Mobility Management for PCS Networks". in *INFOCOM (3)*, pp. 177-1384, 1999.
- [11] F.Bai and A.Helmy, "A survey of Mobility Models in Wireless Ad hoc Networks", Kluwer Academic Publishers, 2004
- [12] M. Musolesi, S.Hailes, C. Mascolo, "An Ad Hoc Mobility Model Founded on Social Network Theory", *Proceedings of MSWiM'04*, pp. 20-24, ACM Press, October, 2004.
- [13] A. Jardosh, E.M. Belding-Royer, K.C.Almeroth, S. Suri. "Towards Realistic Mobility Models For Mobile Ad Hoc Networks", *Proceedings of 9th Annual International Conference on Mobile Computing and Networking (MobiCom 2003)*, San Diego, CA, pp. 217-229, September 2003.
- [14] F. Bai, N. Sadagopan, A.Helmy. "Important: A framework to systematically analyze the Impact of Mobility on Performance of Routing Protocols for ad hoc Networks", *Proceedings of IEEE Information Communications Conference (INFOCOM '03)*, 2003.

- [15] C. Bettstetter, C. Wagner. "The Spatial Node Distribution of the Random Waypoint Mobility Model", *Proc. German Workshop on Mobile Ad-Hoc Networks (WMAN)*, Ulm, Germany, GI Lecture Notes in Informatics, no.P11, pp.41-58, Mar 25-26,2002.
- [16] C. Bettstetter, "Smooth is better than Sharp: A Random Mobility Model for Simulation of Wireless Networks", *Proceedings of the 4th ACM International Workshop on Modeling, Analysis, and Simulation of Wireless and Mobile Systems(MSWiM'01)*, Rome,Italy, July 2001.
- [17] DEM, <http://data.geocomm.com/dem/>
- [18] TIN, http://www.ianko.com/resources/triangulated_irregular_network.htm
- [19] VRML, www.w3.org/Markup/VRML
- [20] Landserf, www.landserf.org
- [21] A* algorithm, www.policymanac.org/games/aStarTutorial
- [22] Mobireal network simulator, <http://www.mobireal.net/>

Simulating a Two-Stage Packet Scheduler

Anton Kos, Sašo Tomažič

University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

Abstract—At present, end-to-end Quality of Service assurance in packet networks is still not appropriately solved. Our contribution to the solution in this area is the proposed merging of the functionalities of the Integrated and Differentiated Services through a selective and combined usage of their advantages. We propose a selective, one-stage or two-stage traffic processing in network devices, supported by a two-stage packet scheduler. When we tried to prove our concept of Quality of Service assurance, we faced the problem of non-existing tools. To facilitate the testing and simulation of the newly developed features, particularly a new packet scheduler, we have developed a simulator for a general model of a network device. The simulator includes modules for its most important elements and functions. We have extensively tested and validated the operation of the simulator with the analytically verifiable settings. Encouraged with the good validation results, we have then simulated our newly developed two-stage packet scheduler. Some of the more interesting simulation results are presented in this article, many more you can find in corresponding references.

Keywords—Deficit Round-Robin (DRR), strict-priority queuing, Differentiated and Integrated Services (DIS), two-stage packet scheduler, simulation.

I. INTRODUCTION

Present connectionless packet networks do not excel at offering end-to-end Quality of Service (QoS) assurance. To overcome their insufficiency in this area of functionality, different solutions have been proposed. While some of the solutions offer absolute flow by flow assurances, they are too complex and do not scale well. Others are simple and scalable, but offer only relative class by class assurances. In our research work we are proposing one possible solution to this problem.

In the following sections we first present the motivation for this work. Then we briefly describe our solution along with the implemented two-stage packet scheduler. Next we briefly explain the operation of the developed simulator with its testing and validation scenarios. At the end we present some interesting simulation results for the two-stage packet scheduler.

II. MOTIVATION

While investigating the possibilities of QoS assurance in packet networks, we have studied the operation of different elements and functions of network devices. A major role in the QoS assurance in a network device is played by the scheduler. We have focused on its operation, particularly on the role of the service discipline used. The choice of service disciplines for packet network schedulers is plentiful. To achieve some sort of QoS assurance, a scheduler should use the appropriate service disciplines. In the next section we propose a new concept of QoS assurance in packet networks. For the purpose of the later we have designed a new scheduler, which uses a combination of two different service disciplines that work on top of each other.

We have found out, that for many service disciplines the analytical results are plentiful, but the simulation results are scarcer. Simulations are on one hand carried out by simulators written in non-dedicated simulation languages (like C++ and similar) resulting in non-reusable code, and on the other hand by complex and powerful simulation tools and packages (like NS and similar) where adding a new functionality and gaining quality simulation results is difficult, extremely expensive or even impossible.

Based on the above we decided to develop a simple general packet network device simulator with all the necessary elements and functionalities. Since it is written in a dedicated high-level object oriented simulation language, it is easy to understand, upgrade and modify. Its modularity makes, for instance changing a service discipline, an easy task. Since it is a network device simulator, it does not include functionalities of the higher protocol levels like TCP or similar. We have also built a number of modules for different schedulers and service disciplines.

III. A TWO-STAGE PACKET SCHEDULER

During our research of QoS assurance in packet networks, we have identified *Integrated Services (IS)* [7] and *Differentiated Services (DS)* [8]-[9] as the two most widely used and promising solutions. A more detailed study and comparison of both has identified their advantages and disadvantages [6] that are almost complementary. We concluded that the best solution is to merge the two concepts into a new solution called a network with *Differentiated and Integrated Services (DIS)* [1], [6].

In a DIS network the type of traffic is recognized and processed accordingly to its QoS demands. Some packets are forwarded without detailed examination, other packets are examined in greater detail and then forwarded, yet still other packets are examined in greater detail, changed accordingly to some rules and then forwarded.

The operation of DIS network is easiest presented through an example. Let us define 6 service classes denoted by letters A to F. In a DS network each of these service classes would be represented by a certain value in the DS field of the IP header. Let us assume that packets from classes D to F do not have high priority. They are placed in the appropriate queue based only on their service class, without further processing. And let us assume that those packets from classes A to C have high priority. They are examined in detail, processed and put into appropriate queue. The class is determined by the value in DS field.

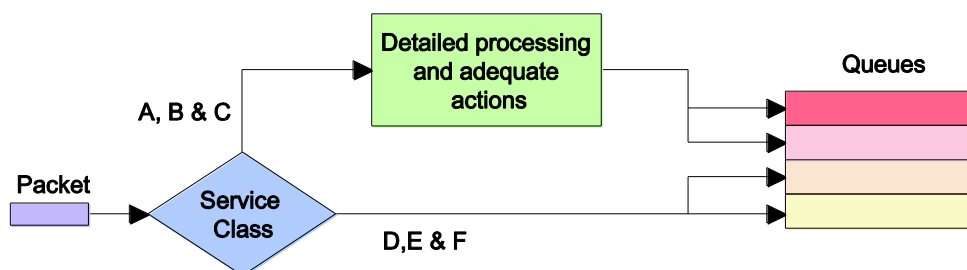


Fig. 1 A simplified packet processing flow in a proposed DIS network.

It can be said that packets of classes D to F receive only one-stage processing, based only on their DS field value, while packets from classes A to C receive two-stage processing. The first stage is based on their DS field value, where it is determined that they need more detailed processing on the second stage. A simplified processing flow for this example is depicted in figure 1.

At the first glance, our proposed DIS network would work like a DS network, meaning that the traffic differentiation would be based on traffic classes. This mechanism implements relative levels of QoS. Since it is anticipated, that for some time the majority of the traffic on public IP network would require only relative QoS, the advantages of DS could be exploited. But for the minority of traffic, with more complex and more stringent QoS demands, a few more service classes are defined. Those classes receive better service with more processing, similar to IS. In addition to higher demands, such traffic would also use more network resources and would therefore exploit advantages of IS concept. Since it is anticipated that it will only be a small part of the total traffic, disadvantages of IS should have only minor effect. At the same time (at least for this traffic) disadvantages of DS would also be eliminated.

At this point it should be stressed, that this is not some sort of "*IS over DS*", but merging of both concepts. In the first stage the network works similar to a DS network. But in the second stage, only for the traffic of certain service classes, the network works similar to an IS network. For more details see [1], [6].

The most critical part of QoS assurance in packet networks and in network devices are schedulers. In general, a scheduler should satisfy the following, sometimes contradictory, demands: simple implementation, low complexity, fairness and flow protection, operation inside agreed boundaries.

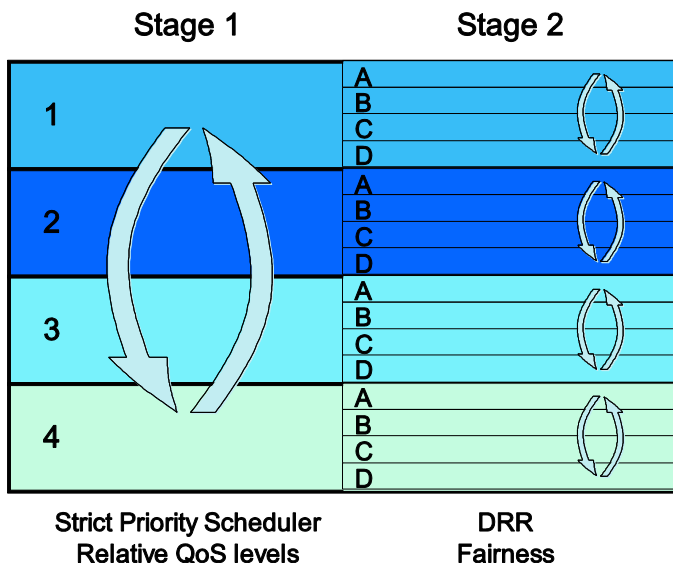


Fig. 2 Output queue configuration in the proposed two-stage packet scheduler for DIS networks.

Since our DIS network introduces two-stage packet processing, it is straightforward to configure output queues in the same manner. This also implies two-stage scheduling (figure 2). The first scheduling stage should assure different levels of QoS between existing service classes, and the second stage should, when necessary, assure fairness and isolation between flows within the same service class. Based on the above conditions, we propose the use of Strict Priority service discipline on the first stage and the use of *Deficit Round Robin* (DRR) service discipline on the second stage. The justification for this is:

- Both schedulers have complexity of $O(1)$.
- **Strict Priority scheduling** assures relative QoS levels between service classes. Referring to the example above, class A has the highest and class F the lowest priority. According to the properties of Strict Priority scheduling, class A packets see the entire link bandwidth, class B packets essentially share bandwidth only with class A packets, and so on till class F packets. Since Class A and B traffic must be limited, it should not happen that their packets would use the entire link bandwidth and in that way starve lower class packets.
- **DRR scheduling** [4] assures fairness and isolation between flows within each service class, meaning that in long term none of the active flows can get more resources than their reservation. Second stage scheduling is reasonable only for high priority service classes like A and B in our example.

The detailed analytical and simulation analysis of two-stage network operation with two-stage scheduling is presented in [1]. Some of the interesting results are also presented later in this paper.

IV. SIMULATOR

For the development of the general packet network device simulator we have used a high level simulation language MODSIM III. It is a general-purpose, modular, strongly typed, block structured simulation language, which provides support for object-oriented programming, discrete event simulation and animated graphics. It is intended to be used for building process-based discrete event simulation models through modular and object-oriented development techniques.

The simulator is built of several modules. Each of the modules includes definitions of one or more simulator objects. Each element or function of a network device is implemented as a separate object. The most important network device objects are: admissioner object, classifier object, queue object, and particularly for this paper the scheduler object. Depending on the settings of a simulation we can have one or multiple instances of the mentioned objects.

A typical example of a simple simulation scenario would be the following. In a single network device we have multiple input ports and multiple output ports (we monitor one output port). On a monitored output port we have a classifier, multiple queues and a scheduler. Each input port has a dedicated traffic generator with its own traffic pattern and generates one traffic flow.

From a packet's point of view the simulator works like this: when a packet object is generated by one of the traffic generator objects, it is sent to one of the input ports of the network device. It is first processed by admissioner. If admitted, it is sent to the classifier on the

appropriate output port. It is then classified into one of the multiple output queues on that output port. Queues are served by the scheduler according to the chosen service discipline.

V. SIMULATOR TESTING AND VALIDATION

We have tested and validated the operation of the simulator with the analytically verifiable settings. First we compared the simulation and analytical results for the simplest queuing system $M/M/1$ [1]-[2]. The results of the simulation are practically the same and almost indistinguishable from the analytical results. With this we have in the first place confirmed the correct operation of traffic generators, the correctness of procedures for choosing simulation parameters, and the correctness of their calculations. Since the $M/M/1$ queuing system is implemented as one FCFS (First Come First Serve) queue, it is hard to say that this also confirms the correct operation of the entire simulator.

To further test and validate the operation of the simulator we have tested a more complex queuing system $M/G/1$ with priorities. We have done so for different values of the system load ρ and traffic priorities p , where $p = 4$ is the highest priority and $p = 1$ is the lowest priority (see details in [1] and [2]).

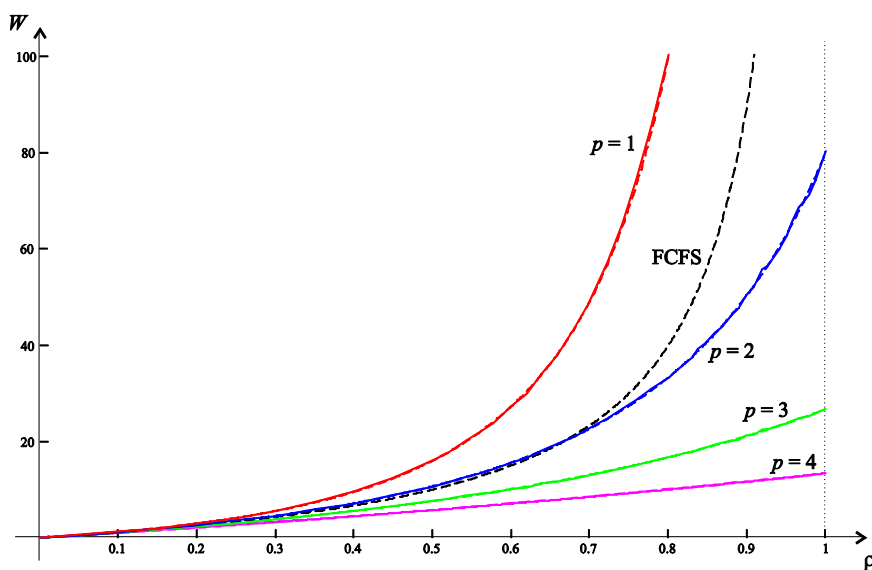


Fig. 3 The comparison of analytical and simulation results in the $M/G/1$ queuing system with priorities p for the average waiting time W at system load $0 < \rho < 1$.

In figure 3 we see the results for the average packet waiting time W . As in previous example, also here the simulation and analytical results are practically the same. That confirms the correct operation of all the elements of our simulator.

We have done similar validation test for the DRR scheduler, two-stage processing and queuing. More on these validations can be found in [1].

VI. SIMULATION RESULTS FOR THE TWO-STAGE PACKET SCHEDULER

Encouraged by the simulator validation results, we have then simulated our two-stage packet scheduler with Strict Priority service discipline on the first stage, and DRR service discipline on the second stage. We have carried out extensive simulations with many different scenarios, which included many different settings. In this article we present our results only briefly, more you can find in [1]. Let us mention, that all figures also include the analytical result for the FCFS scheduler, to which we compare all simulation results.

We have simulated a modified version of DRR scheduler. The operation of an original DRR scheduler is defined in [4]; our modifications are explained in detail in [1]. The main characteristic of all Deficit Round Robin (DRR) like scheduling algorithms is their ability to provide guaranteed service rates for each flow (providing that each flow has its own queue).

DRR services flows in a strict round-robin order. It has complexity $O(1)$ and it is easy to implement. Its latency is comparable to other frame-based schedulers. A detailed operation of DRR algorithm can be found in [4]. Below is the list of variables used:

- R transmission rate of an output link,
- N the total number of active flows,
- r_i the reserved rate of flow i ,
- w_i weight assigned to each flow i ,
- Q_i quantum assigned to flow i .

Because all flows share the same output link, a necessary constraint is that the sum of all reserved rates must be less or equal to the transmission rate of the output link:

$$\sum_i r_i \leq R \quad (1)$$

Let r_{min} be the smallest of all r_i . Each flow i is assigned a weight that is given by:

$$w_i = \frac{r_i}{r_{min}} \quad (2)$$

Note that for all i in $\{1, 2, \dots, N\}$ holds $w_i \geq 1$. Each flow i is assigned a quantum of Q_i bits that is a whole positive value. This quantum is actually the amount of service that the flow should receive during each round robin service opportunity. Let us define with Q_{min} the minimum of all the quanta. Then the quantum for each flow i is expressed as:

$$Q_i = w_i Q_{min} \quad (3)$$

The consequence of the above constraints and definitions is, that if a flow i sends more data as it is entitled to through the size of its quantum Q_i (misbehaved flow), its queue will become longer and its packets will experience greater delay, while the other flows will remain unaffected. This implies that DRR is fair and provides flow isolation.

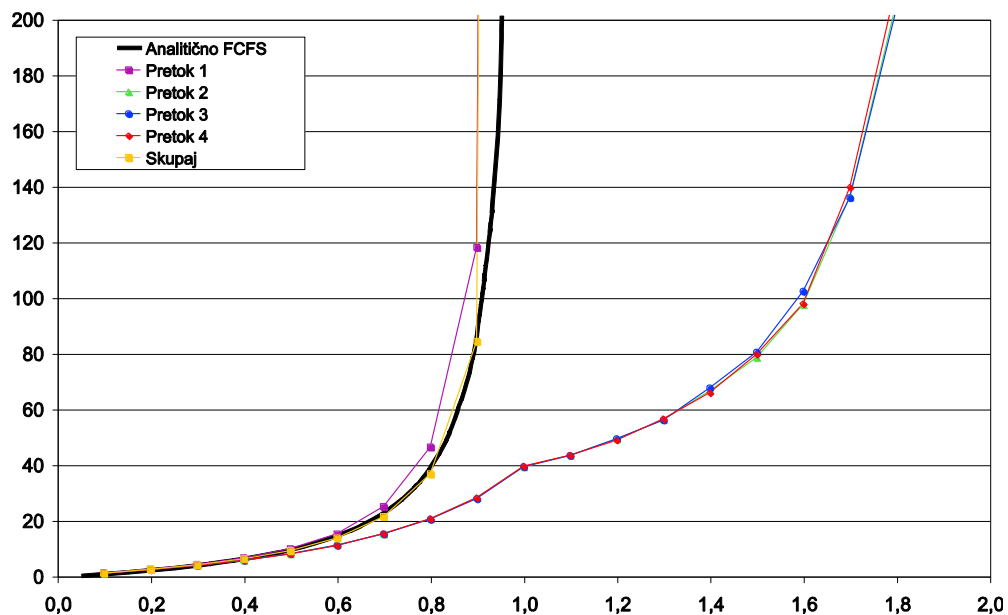


Fig. 4 Average packet delays of DRR scheduler at $Q_i = 2000$ bytes, at exponential packet length distribution. Flow 1 is misbehaved, sending 5 times more data than agreed.

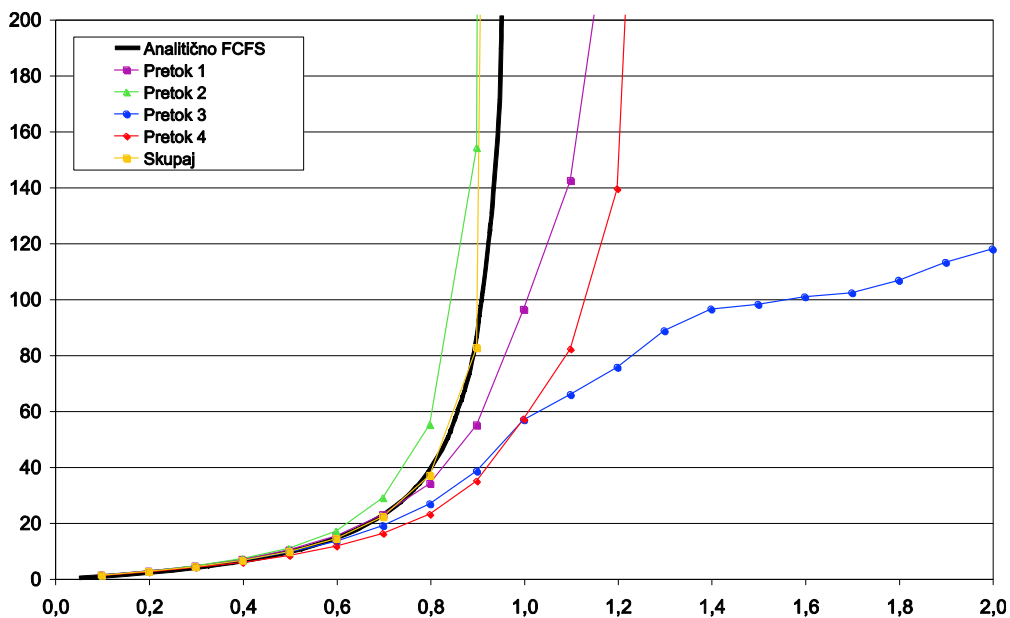


Fig. 5 Average packet delays of DRR scheduler at $Q_1 = Q_2 = 3000$ and $Q_3 = Q_4 = 9000$ bytes, at exponential packet length distribution. Flow 2 is misbehaving.

To prove the above claims, we have first simulated a scenario where all the traffic has the same priority, as if there was no first stage in the two-stage scheduler. Then we set traffic flow 1 to be misbehaving by sending 5 times more data than it is entitled to through its quantum. In figure 4 we see that the delay of the misbehaving flow 1 quickly rises above all boundaries as the system load ρ approaches 1. When ρ exceeds 1 (overload that is caused by the misbehaving flow) it only affects flow 1, the other three flows are unaffected.

In figure 5 we see another scenario with average delays of flows that have quantum values $Q_1 = Q_2 = 3000$ and $Q_3 = Q_4 = 9000$ bytes, and exponential packet length distribution. Flow 2 with reservation $Q_2 = 3000$ bytes is misbehaving (sending more data than agreed). The results show that its delays very quickly rise above all limits while delays of other flows, which behave as agreed, experience expected delays.

We gain very similar results for the two-stage scheduler. We use the Strict Priority service discipline on the first stage and the DRR service discipline on the second stage. In figure 6 all flows have the same quantum. We see that flows are differentiated by its delay according to the priority class they belong to. But flows inside the same priority class have exactly the same average delay.

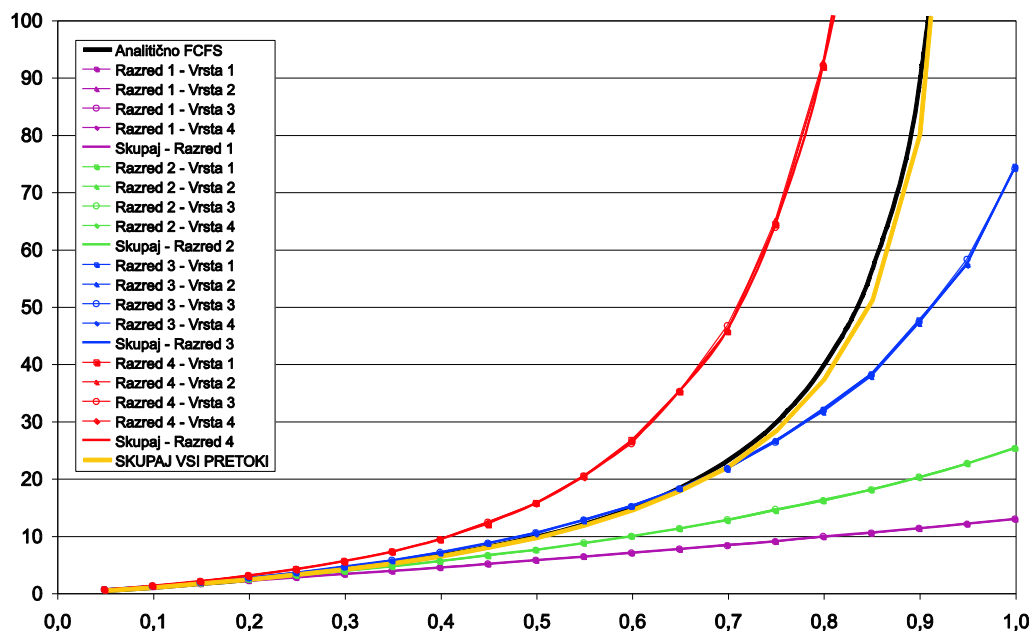


Fig. 6 Average packet delays of two-stage scheduler with strict priority scheduler on the first stage and DRR scheduler on the second stage. The quantum for all the flows is the same at $Q_i = 3000$.

Situation is a bit different at the presence of a misbehaving flow. In figure 7 we have one misbehaving flow in each of the priority classes. We see that the delay increases only for the misbehaving flows, and that other flows inside the same priority class are not affected. But since we have strict priority queuing on the first stage, misbehaving flows of high priority classes affect all the flows in low priority classes.

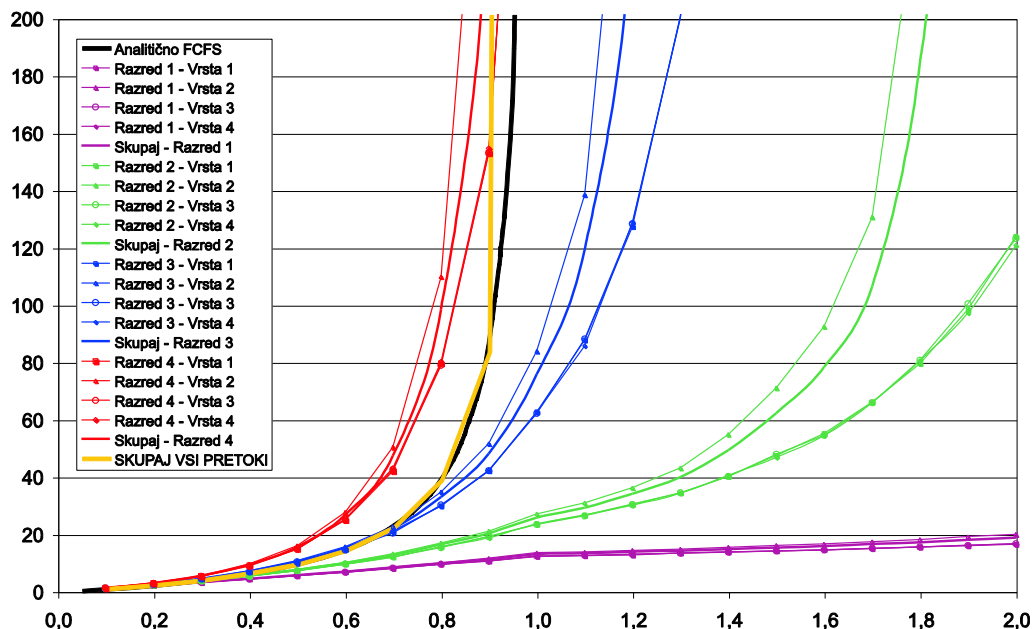


Fig. 7 Average packet delays of two-stage scheduler with strict-priority scheduler on the first stage and DRR scheduler on the second stage. The quantum for all the flows is $Q_i = 3000$.

Flow 2 in each of the priority classes is misbehaving.

The simulation results show, that the operation of our two-stage packet scheduler is up to the expectations. High priority traffic classes experience low delay and are not affected by the traffic of lower priority classes. Of course, the opposite is true for the low priority classes. An important improvement over the classic one-stage Strict Priority service discipline is the DRR service discipline on the second stage. Its main contribution is the ability to provide fairness and isolation among flows of the same traffic class.

VII. CONCLUSION

We propose a new QoS assurance concept in packet networks and consequently a new two-stage packet scheduler. We developed and tested a network device simulator, which has proven its functionality and provided us with interesting results for the two-stage packet scheduler. Since the simulator is built-up of modules it can be easily upgraded. We can easily add new functionalities, for instance write a new service discipline, or reuse some of its modules in other simulators.

The simulation results have shown, that with the choice of the two very simple, undemanding and easily implementable service disciplines, we can exploit the advantages of both and at the same time suppress most of their disadvantages. In our two-stage packet scheduler the first stage with Strict Priority service discipline provided traffic class differentiation based on delay, while the second stage with DRR service discipline provided fairness and isolation of the traffic flows within those classes. The modular design of the simulator allows for a lot of further work to be done with not too much effort.

REFERENCES

- [1] Anton Kos, "Zagotavljanje različnih stopenj kakovosti storitev v omrežjih s paketnim prenosom podatkov", Ph.D. Thesis, Faculty of Electrical Engineering, University of Ljubljana, 2006.
- [2] Leonard Kleinrock, "Queueing Systems, Volume I: Theory", John Wiley & Sons, 1975
- [3] Robert Verlič, Anton Kos, Sašo Tomažič, "Zagotavljanje kakovosti storitev v paketnih omrežjih", Elektrotehniški vestnik, vol. 73(2-3), 2006.
- [4] M. Shreedhar, George Varghese, "Efficient Fair Queuing Using Deficit Round Robin", IEEE/ACM Transactions on Networking, Volume 4, Issue 3, June 1996.
- [5] Anton Kos, Jelena Miletic, Sašo Tomažič, "On Latency of Deficit Round Robin", ERK 2006, Zvezek A, strani 171-174, Portorož, september 2006.
- [6] Anton Kos, Sašo Tomažič, "Merging of Integrated and Differentiated Services", VIPSI-2006 Montreal, Montreal, Canada, 2006.
- [7] RFC 1633, "Integrated Services in the Internet Architecture: an Overview", IETF, July 1994
- [8] RFC 2474, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", IETF, December 1998.
- [9] RFC 2475, "An Architecture for Differentiated Services", IETF, December 1998.