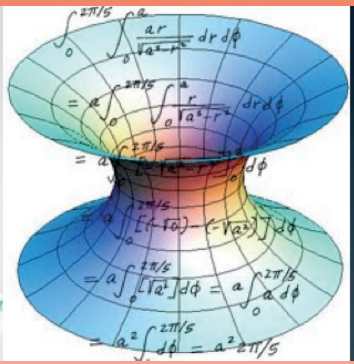
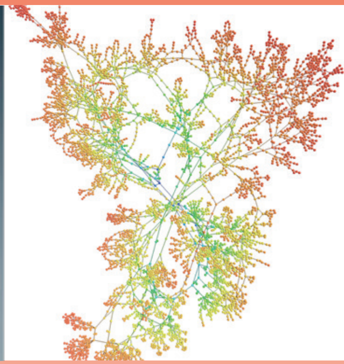
3RD URV DOCTORAL WORKSHOP IN COMPUTER SCIENCE AND MATHEMATICS

Edited by Sergio Gómez & Aïda Valls Mateu





UNIVERSITAT ROVIRA i VIRGILI

[DΣIM] Departament d'Enginyeria
Informàtica i Matemàtiques

Escola Tècnica Superior
d'Enginyeria
Universitat Rovira i Virgili

Title: 3rd URV Doctoral Workshop in Computer Science and Mathematics
Editors: Sergio Gómez & Aïda Valls Mateu
November 2016

Universitat Rovira i Virgili
C/ de l'Escorxador, s/n
43003 – Tarragona
Catalunya (Spain)

ISBN: 978-84-8424-495-0

Preface

This book of proceedings gathers the contributions presented at the *3rd URV Doctoral Workshop in Computer Science and Mathematics*. After the successful previous editions in 2014 and 2015, the third edition has been held in Tarragona (Catalonia, Spain) on November 17th, 2016. It has been jointly organized by the research group Algorithms Embedded in Physical Systems (ALEPHSYS) and the Doctoral Program on Computer Science and Mathematics of Security of Universitat Rovira i Virgili (URV). The main aim of this workshop is to promote the dissemination of the ideas, methods and results that are developed in the Doctoral Thesis of the students of this doctorate program, and to promote the knowledge, collaboration and discussion between their respective research groups.

The workshop had two invited talks and twelve oral presentations. The first invited talk was given Prof. Jordi García-Ojalvo, the leader of the Dynamical Systems Biology Lab at Universitat Pompeu Fabra, who talked about how models can be applied to understand life, from gene circuits to neural networks. The second invited talk was given by Prof. Jordi Vitrià, coordinator of the BCN Perceptual Computing Lab at Universitat de Barcelona. He provided an overview and practical hints of one of the currently most important topics in Artificial Intelligence: Deep Learning.

In this book, the reader will find the contributions of the Ph.D. students. Each chapter presents the research topic of one student, the goals and some of the results. It is worth to note the wide coverage of this workshop, with contributions to the following main research lines: (1) Security and privacy in computer systems, (2) Artificial intelligence, robotics and vision, (3) Telematic architectures and complex networks, and (4) Mathematics. All contributions present innovative proposals, methods or applications, with the aim of opening new and strategic research lines.

The editors and organizers invite you to contact the authors for more detailed explanations and we encourage you to send them your suggestions and comments that may certainly help them in the next steps of their PhD thesis. The organizing committee was formed by Dr. Sergio Gómez, Dr. Aïda Valls

(Coordinator of the Ph.D. program), Mr. Joan T. Matamalas and Mrs. Olga Segú.

We could not finish without first thanking the invited speakers for accepting to contribute and for giving us such interesting conferences. Second, we thank all the participants and, especially, the students that presented their work in this DCSM workshop. Finally, we also want to thank Universitat Rovira i Virgili (URV), the Departament d'Enginyeria Informàtica i Matemàtiques (DEIM), and the Escola Tècnica Superior d'Enginyeria (ETSE) for their support.

Sergio Gómez and Aïda Valls (Editors)

Contents

The local metric dimension of the lexicographic product of graphs
Gabriel A. Barragán-Ramírez 3

Mobility analysis and prediction for Smart Health and Beyond
Abdulrahman Qasem Al-Molegi 7

Real-Time Traffic-Scheduling in Underwater Acoustic Wireless Sensor Networks
Pere Millán 11

Making Decisions with Semantic Criteria
Miriam Martínez-García 17

Efficient Wireless Channel Characterization for Context-Aware Scenarios
Fran Casino 21

Reducing Network Overhead on Personal Cloud Systems
Raúl Sáiz-Laudó 27

Privacy-Preserving Statistical Computation in the Cloud
Sara Ricci 33

A Numerical Study on Bayesian Search Strategies
Lluís Arola Fernández 39

Clinical Decision Support System for Diabetic Retinopathy Risk Evaluation
Emran Saleh 45

Searchable Encryption for Geo-Referenced Data
Jordi Ribes González 51

On semi-automatic methods for localization and segmentation of optic disk in digital fundus images
José Escorcia-Gutierrez 55

Classification of computer technologies for multimorbidity
Wilfrido Ortega León 59

The local metric dimension of the lexicographic product of graphs

Gabriel A. Barragán-Ramírez

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
gabrielantonio.barragan@estudiants.urv.cat

1 Introduction

In a metric space (M, d) we say that an element $a \in M$ distinguishes other two $b, c \in M$ if $d(a, b) \neq d(a, c)$. $B \subseteq M$ is a metric generator for M if every pair of elements in M is distinguished by some element in B. The metric dimension of the space is defined as the minimum number that can be cardinality of a generator set. The notion of metric dimension was introduced by Blumenthal in [1] in 1953 and in the context of graph theory independently by Harary and Melter [2] and Slater [3] in 1975. We can define a metric on an graph considering the distance between two vertices as the length of the shortest path between them. In 1996 Khuller et al. [4] proved that the decision problem of determining whether the metric dimension is less than a given value, is NP-complete. My thesis is focused in a variant of the metric dimension, the *Local Metric Dimension* introduced by Okamoto et al. in 2010 [5]. When we calculate the local metric dimension of a graph we are concerned only to distinguish pairs of adjacent vertices. The calculation of the local metric dimension is also proved to be a NP-Hard problem by Rodríguez-Velázquez and Fernau in 2014 [6]. The aim of the thesis is to tackle the problem of calculating the local metric dimension of graphs that are obtained from simpler ones by graph operations called in general graph products. For a detailed study of such products we refer to [9].

2 Basic definitions and results

Let $G(V, E)$ be a graph, non oriented and without multiple edges. For a pair of vertices $u, v \in V$ we denote by $d_G(u, v)$ the length of a shortest u, v-path. It is clear that (G, d_G) is a metric space. $B \subseteq V$ is a local metric generator for G if for every $uv \in E$ there exists $x \in B$ such that $d_G(x, u) \neq d_G(x, v)$. The minimum r such that r can be the cardinality of a local metric generator is

[⋆] PhD advisor: J. A. Rodíguez-Velázquez called the *local metric dimension* of G that is denoted by $\dim_l(G)$ and a local metric generator with that cardinal is called a *local metric basis* for G.

Okamoto's et. al. article [5] is rich in results. Among them

Theorem 1. *If $G\Box H$ denotes the Cartesian product of the graphs G and H then*

$$\dim_l(G\Box H) = \max\{\dim_l(G), \dim_l(H)\}.$$

That is the origin of our research. We study the local metric dimension in the *strong product*, the *leixcographic product*, the *corona product* of graphs, in the graphs obtained by *point attaching* and also the *simultaneous local metric dimension* of families of lexicographic product graphs.

As an illustration of our work we will focus on our last article that is about the calculation of the local metric dimension of lexicographic product graphs [10].

2.1 The lexicographic product

Let G be a graph of order n, and let $\mathcal{H} = \{H_1, H_2, \dots, H_n\}$ be an ordered family composed by n graphs. The *lexicographic product* of G and $\mathcal{H}$ is the graph $G \circ \mathcal{H}$, such that $V(G \circ \mathcal{H}) = \bigcup_{u_i \in V(G)}(\{u_i\} \times V(H_i))$ and $(u_i, v_r)(u_j, v_s) \in E(G \circ \mathcal{H})$ if and only if $u_iu_j \in E(G)$ or $i = j$ and $v_rv_s \in E(H_i)$.

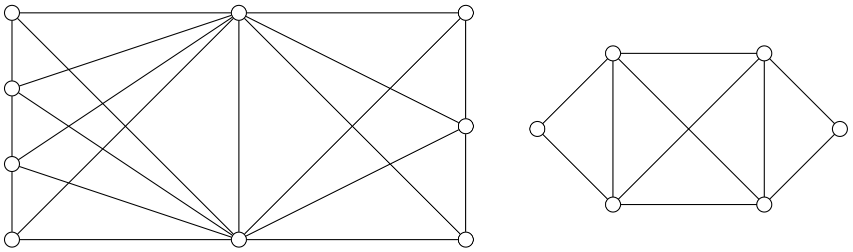


Fig. 1: The lexicographic product graphs $P_3 \circ \{P_4, K_2, P_3\}$ and $P_4 \circ \{H_1, H_2, H_3, H_4\}$, where $H_1 \cong H_4 \cong K_1$ and $H_2 \cong H_3 \cong K_2$.

If G is a connected graph and (u_i, b) and (u_j, d) are vertices of $G \circ \mathcal{H}$, then

$$d_{G\circ\mathcal{H}}((u_i, b), (u_j, d)) = \begin{cases} d_G(u_i, u_j), \text{ if } i \neq j, \\ \\ d_{H_i,2}(b, d), \text{ if } i = j. \end{cases}$$

Where $d_{H_i,2}(b, d) = \min\{d_H(b, d), 2\}$ is the *two distance* in the graph H_i. For each $H_i \in \mathcal{H}$, $(H_i, d_{H_i,2})$ is a metric space and makes sense the question of the (local) metric dimension in this space. This parameter is called *(local)* *adjacency dimension* of H_i and it is denoted by $\operatorname{adim}(\operatorname{adim}_l)(H_i)$. The adjacency dimension was introduced by Jannesari and Omoomi in [8] as a tool to study the metric dimension of lexicographic product graphs. The study of the local version of the adjacency dimension was introduced by Fernau and Rodríguez-Velázquez in [7] where they calculate the meric dimension and the local metric dimension of the corona product of graphs.Also in this paper they prove that the problem of computing the (local) adjacency dimension is NP-hard.

Theorem 2. *Let G be a connected graph of order $n \geq 2$, let $\{U_1, U_2, \ldots, U_k\}$ be the set of non-singleton true twin equivalence classes of G and let $\mathcal{H} = \{H_1, \ldots, H_n\}$ be a family of graphs. Then*

$$\dim_l(G \circ \mathcal{H}) = \sum_{i=1}^{n} \operatorname{adim}_l(H_i) + \sum_{I \cap U_j \neq \emptyset} (|I \cap U_j| - 1) + \varrho(G, \mathcal{H}).$$

The parameter $\varrho(G, \mathcal{H})$ is well defined and we have worked on conditions to it be equal zero, for example in the case $N_r \notin \mathcal{H}$.

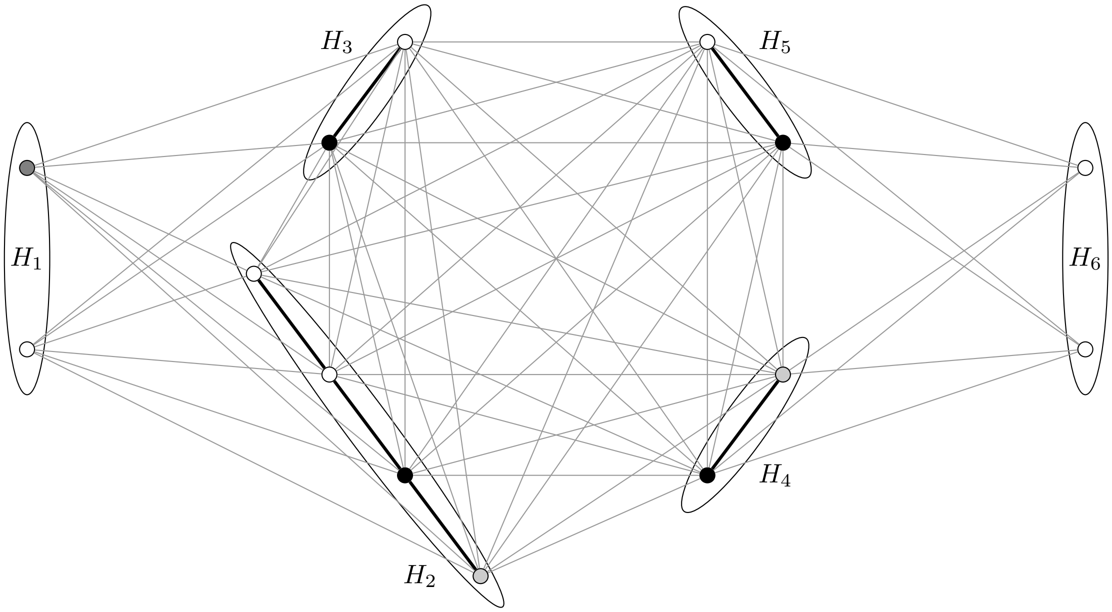


Fig. 2: The graph $G \circ \mathcal{H}$, where G is the right-hand graph shown in Figure 1 and $\mathcal{H}$ is the family composed by the graphs $H_1 \cong H_6 \cong N_2$, $H_2 \cong P_4$, $H_3 \cong H_4 \cong H_5 \cong K_2$. The black vertices correspond to the local adjacency basis, the light grey to the vertex cover of the twins graph and the dark-grey vertex stands for the ϱ parameter. The set of black- and grey-coloured vertices is a local metric basis of $G \circ \mathcal{H}$.

References

[1] L.M. Blumenthal Theory and Applications of Distance Geometry. Clarendon Press, Oxford (1953)

[2] F. Harary, R.A. Melter On the metric dimension of a graph. *Ars Comb.*, 2 (1976), pp. 191–195

[3] P. J. Slater. Leaves of trees. Proc. 6th Southeastern Conference on Combinatorics, Graph Theory, and Computing (Florida Atlantic Univ., Boca Raton, Fla., 1975). Congressus Numerantium 14, Winnipeg: Utilitas Math., pp. 549–559

[4] S. Khuller, B. Raghavachari, A. Rosenfeld. Landmarks is Graphs. *Discrete Applied Mathematics* 70 (3) (1996) 217-229

[5] F. Okamoto, B. Phinezy, P. Zhang. The local metric dimension of a graph. *Mathematica Bohemica* 3, 135 (2010) 239–255

[6] H. Fernau, J. A. Rodríguez-Velázquez. On the (adjacency) metric dimension of corona and strong product graphs and their local variants: combinatorial and computational results. arXiv:1309.2275 [math.CO].

[7] H. Fernau, J. A. Rodríguez-Velázquez. Notions of metric dimension of corona products: combinatorial and computational results. Computer science—theory and applications, vol. 8476 of Lecture Notes in Comput. Sci., Springer, Cham, 2014, pp. 153–166.

[8] M. Jannesari, B. Omoomi. The metric dimension of the lexicographic product of graphs. Discrete Mathematics 312 (22) (2012) 3349–3356.

[9] R. Hammack, W. Imrich, S. Klavžar. Handbook of product graphs, Discrete Mathematics and its Applications. 2nd ed., CRC Press, 2011.

[10] G. A. Barragán-Ramírez, A. Estrada-Moreno, Y. Ramírez-Cruz, J. A. Rodríguez-Velázquez. The local metric dimension of the lexicographic product of graphs. arXiv:1602.07537

Mobility analysis and prediction for Smart Health and Beyond

Abdulrahman Qasem Al-Molegi [⋆]

Smart Health Research Group
Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
abdulrahmanqasemyahya.almolegi@urv.cat

Abstract. In this article we recall the importance of mobility data analysis for the proper prediction of human behaviours. Also, we emphasise its multiple applications to location-based services, recommendation systems, route planners, and so on.

One of the key goals of mobility analysis is to predict regions of interest and the next interest point in which a given person could be found. We summarise the initial steps that we have taken towards proposing a novel method to achieve these goals.

1 Introduction

Human behaviour is very complex and diverse. Mobility, as a component of human behaviour, is also complex, but its variability is lower and could be studied with more focused approaches. In most cases, human mobility is analysed with the goal of predicting future behaviours.

Monitoring people's mobility during their daily activities is a basic requirement to provide advanced location-based services (LBS)[1,2,3]. In this sense, mobility data play a key role in the analysis of people behaviours, including predicting their next location. Fortunately, due to the rapid enhancement of data collection abilities of mobile devices, we can easily collect large amounts of people's mobility data at a very low cost.

There is a variety of devices to collect mobility data. Smartphones are considered one of the most appropriate options for tracking and recording user mobility during daily activities due to their proximity to the users and their ability to carry multiple sensors. The widespread usage of smartphones together with the development of location-based applications and services have received considerable interest, and special attention is paid towards building efficient methods for analysing and predicting important locations, where smartphone users will be next. These methods aim to improve both end-user applications, such as healthcare applications [4], recommendation sys-

[⋆] Ph.D. advisor: Dr. Agusti Solanas

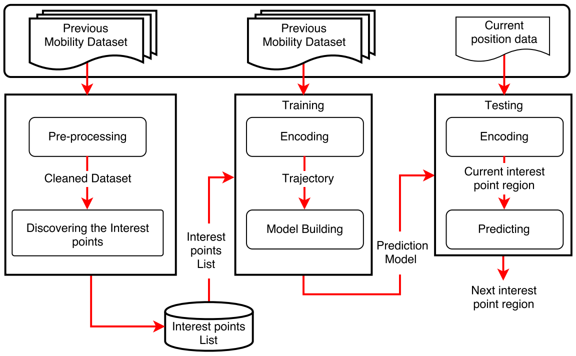


Fig. 1: General architecture for next locations prediction.

tems [5], route planning, carpooling, meeting planners or location-based advertisements, and also to help the corresponding institutions to solve issues related to network management, healthcare, human computer interaction, socio-economic modelling for urban planning, public transportation planning, public safety assurance, etc [6].

Several research papers discussed the problem of discovering interest points and regions of interest and predicting people's next locations based on GPS trajectory data. Those approaches can be roughly classified into: (i) probabilistic models, such as Markov model and (ii) supervised learning models, such as Association Rules, Support Vector Machine and Neural Networks.

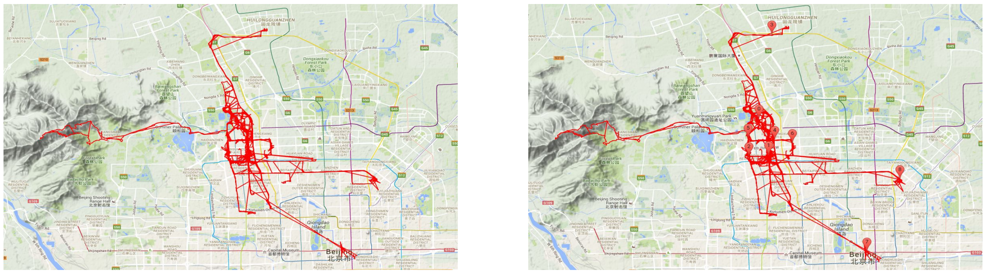
Next, we describe our initial steps into proposing novel methods to predict regions of interest and interest points based on mobility data.

2 General Architecture

Figure 1 shows the main steps involved in the next locations prediction approach that we are currently studying. In the first step, a pre-processing of the dataset is performed to remove possible noise from the data and then to discover the interest points located in the user movement region. The second step corresponds to the building of a prediction model. In the last step, the prediction model is evaluated with testing data.

2.1 Discovering the Regions with Interest Points

In order to build an accurate prediction model, the regions with interest points of the trajectory, that properly describe the movement of the user in the region, must be identified. Many popular algorithms for discovering significant places depend on the extraction of consecutive GPS points from trajectories that satisfy some given threshold conditions (e.g. stay time, distance, etc). If consecutive GPS points satisfy those conditions, it is assumed that a significant place has been discovered (*cf.* Figure 2). It could be observed that, in



(a) User trajectory.

(b) The discovered interest points.

Fig. 2: Discovering the interest points.

general, the density of GPS points in the region of interest points is higher than in others because people tend to either move slowly or don't move at all in those regions.

The algorithm that we are currently studying for discovering interest points must be sound and complete. The algorithm to discover interest points is sound if it only finds interest points, and complete if it finds all interest points. These two properties are used as a metric to evaluate the quality of the algorithm.

2.2 Model building

To predict people's future locations, learning techniques like Markov Models, Association rules, Bayesian Networks or Neural Networks are obvious candidates to be applied. One of the challenges that are faced by researchers while predicting people movements is how to transfer (adapt) these techniques to work with the context information of the movements. Building an accurate prediction model for all users is hard or even impossible because the next location prediction is a user specific problem. Even if the visited locations might overlap among different users the trajectory of user visits different location is most likely unique. Thus, building one prediction model for each user could be desirable. Usually, building the model, i.e. discovering the frequent trajectories and location, is performed off-line while the prediction itself is performed on-line.

We propose to apply Markov Chains to address this prediction problem. The Markov Chain (MC) model is a technique that naturally finds its utility in movement prediction. To build a MC prediction model, the transitions matrix between the interest points region are computed. The rows of the matrix represent the last visited interest point region while the columns represent the next interest point region. If the user never moves between two interest points, that transition value is set to zero.

We extend the common MC model by including the notion of time. The transitions and time matrices between the interest points region are computed. Our time matrix represents the different movement time between locations. We proposed to integrate the transition and time matrices into one matrix, called Tran-Time matrix. The proposed model essentially contains the spatio-temporal properties that have emerged from the GPS coordinates and the associated times.

3 Conclusion

This brief article presents a general framework for the prediction of the next locations of users. Discovering the interest points in the user movement area and then predicting the future context information of people movements are the two main steps in the prediction model. Soundness and completeness are used as metrics to evaluate the proposed algorithms for discovering interest points. We have shown that the MC model can be extended to include the time associated with the GPS points.

We are currently investigating the application of our model to real data and we expect to obtain preliminary results in the next few months.

Acknowledgement. The author is a Martí-Franquès grant-holder at the Universitat Rovira i Virgili, Tarragona - Spain.

References

[1] Rebollo-Monedero, D., Forné, J., Solanas, A., Martinez-Balleste, A., Private location-based information retrieval through user collaboration. *Computer Communications 33(6): 762-774,* 2010.

[2] Solanas, A., and Martinez-Balleste, A TTP-free protocol for location privacy in location-based services. *Computer Communications 31(6): 1181-1191* 2008.

[3] Solanas, A. and Domingo-Ferrer, J. and Martínez-Ballesté, A. Location Privacy in Location-Based Services: Beyond TTP-based Schemes. *PiLBA* 2008.

[4] A. Solanas, C. Patsakis, M. Conti, I. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. Perez-Martinez, R. Di Pietro, D. Perrea, and A. Martinez-Balleste, Smart health: A context-aware health paradigm within smart cities. *IEEE Communications Magazine, Vol. 52, No. 8, pp. 74-81*, August, 2014.

[5] Casino, F., Domingo-Ferrer, J., Patsakis, C., Puig, D., and Solanas, A., A k-anonymous approach to privacy preserving collaborative filtering. *J. Comput. Syst. Sci. 81(6): 1000-1011*, 2015.

[6] Asgari, F. and Gauthier, V. and Becker, M., A survey on Human Mobility and its Applications. *arXiv:1307.0814v1, pp. 1-18*, 2013.

Real-Time Traffic-Scheduling in Underwater Acoustic Wireless Sensor Networks

Pere Millán[*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
pere.millan@urv.cat

Abstract. Underwater sensor networks are an important field of research. Several applications require the use of this kind of networks, like tsunami or oil spill alerts. The underwater medium is very harsh and only acoustic signals can be used for transmitting information. This kind of networks is still in development, far from reaching standard consensus on basic aspects like carrier frequency or modulation techniques. The use of these networks for real-time applications has not been analyzed previously. This paper summarizes [1], where we present two solutions for the scheduling of real-time messages and we provide a time constraint analysis of the network performance.

Keywords: Underwater Sensor Networks, Acoustic Sensor Networks, Environmental Monitoring

1 Introduction

Underwater acoustic wireless sensor networks (WSN) are becoming a hot research topic as they have turned into the primary tool to monitor and act upon the well-being of marine environments. Radio frequency electromagnetic signals do not propagate well underwater. Huge amount of power is required to transmit messages even for short distances. The presence of particles and moving obstacles like fish, prevents the use of optical carriers. For underwater transmissions, the best option are acoustic carriers. While WSN based on RF transmissions have been studied and several protocols have been proposed, the solutions achieved for them are not useful for acoustic underwater networks, since propagation delay is usually larger than transmission time. A message may be received well after its transmission has finished in the source node.

Real-time (RT) communications require not only that messages are transmitted properly, but also before a particular instant named deadline. If the deadline is missed, the message is not valid and may have serious consequences. A feasible RT schedule is one in which all messages comply with

[*] PhD advisors: Carlos Molina (URV), Roc Meseguer (UPC). Additional authors: R. Santos, J. Orozco, M. Micheletto, S.F. Ochoa.

their deadlines. RT message scheduling in multi-hop networks is a complex problem that requires the use of routing and queueing techniques. If all the nodes in the network have a direct link to the rest of the nodes, the problem may be solved using an integer linear programming approach. However, when a message should go through intermediate nodes, it is not only a question of when a node should transmit (MAC problem) but also of selecting the appropriate path. In this case, the shortest path is not always the best one, as a per-node scheduling should be incorporated in the analysis. In fact, a node holding more than one message has to schedule their transmission introducing additional delays.

In this paper we extend the proposed algorithm presented in [2] to include RT constraints and message transfers between any pair of nodes in the system. A TDMA (Time Division Multiple Access) access protocol is proposed with an off-line allocation and scheduling algorithm. Feasibility conditions are given for the system to operate with hard RT constraints.

2 System Model

For the sake of simplicity, we assume the propagation delay between two nodes within transmission range is equal in both directions. Any node may transmit a message to any other node in the network if there is a valid path between both of them. We denote a message from node a to node b as m_{ab}. We also assume all messages require one time slot to be transmitted and that they are sent periodically. Additionally, all messages should be received before the associated deadline. P_{ab} and D_{ab} represent the period and deadline respectively. In general we define $Z = \{m_{ij}(P_{ij}, D_{ij})\}$.

The network can be modeled as a directed graph $G = (V, E)$, in which V is the set of nodes in the network and E the set of edges. If two nodes u and v are within transmission range, there is an edge connecting them, $e = (u, v)$. Each edge has a label that represents the transmission delay between the nodes measured in time slots, τ_{uv}. As collisions are important only if they are produced at the node, there are four different scenarios, as stated in [3].

We propose a slot allocation method to order the access of the nodes to the channel in such a way that each message originated in a node may reach its destination node without collisions. We begin considering that destination nodes are within transmission range of source/transmission node, and later we extend the analysis for nodes at larger distances. Stated in this way, the slot assignment problem is an extension of the graph coloring problem. We present an integer linear programming (ILP) model, to minimize the frame length measured in slots. The model is significantly more complex if a per message slot allocation is performed. Further details can be found in [3].

3 Scheduling

Path discovery is a well known problem in networking. Several algorithms have been proposed to compute the best path for a message to reach destination from a source. The most common solutions are based on Dijsktra algorithm to determine the shortest path from any node in the network to any other node (SPF, shortest path first). In the case of communication networks, the cost associated to the edges may be related to the actual delay between the nodes, an economical cost for using that link (paying service to a third party company) or the power required to use the link. For real-time messages, the total delay in the path should be less or equal to the deadline of the message. If this condition is not guaranteed, the message is not schedulable and the network does not comply the real-time requirements.

4 Heuristic approach

In the proposed model, the variables that affect the communication speed and therefore the timing of the system are the frame duration, the order of transmissions and reception of messages, and the routes that each message follows within the network. In section 5 of [1] an heuristic algorithm is presented to optimize the message/slot allocation to minimize the frame size and guarantee the deadlines. The minimum length frame is not necessarily the optimal to meet the system time requirements and this impedes uncoupling the calculation of the frame with respect to the calculation of routes. The heuristic presented generates a fixed length frame and optimize the paths of the messages to meet all system deadlines. This heuristic is better suited for complex problems where exists multiple paths for different messages.

5 Real-time applications

Tsunamis are generated by earthquakes in the ocean, and can be tragic like the ones in Japan 2011 or Indonesia 2004. While in the case of Japan, the number of casualties associated to the tsunami is relatively low, in the case of Indonesia, the number of victims is counted in thousands (and severe economic loses in infrastructure). The difference in the number of victims is associated to the early alert that people in Japan received to get into a safe place.

Detecting a tsunami is a hard work. Seismic sensors may be deployed in the area in which the earthquake may take place (geologic fault) and if this is detected, depending on the intensity a tsunami alert may be issued. The time available between the earthquake and the arrival of the wave to the beach depends on the distance to the earthquake epicenter. However, it is clear that there is a hard RT restriction as the alert should be issued with enough time for people to move into a safe place.

The system may have some buoys anchored along the fault and linked to the seismic sensors so once the earthquake is detected, the buoy connects through a satellite network to a management disaster office reporting the event, intensity and tsunami probability. However, buoys are vandalized by pirates or even fishermen jeopardizing the network operation. To avoid this, an underwater acoustic WSN is proposed operating in RT. The network deployment, nodes distribution and number of hops discussion is out of the scope of this paper. However, the RT analysis and network performance modeling proposed here can be used to set-up the appropriate network.

6 Conclusions and Future Work

We presented a RT analysis for an underwater acoustic WSN, with two approaches. First, the network is analyzed with integer linear programming techniques. SPF is used as routing policy combined with a message or node slot allocation procedure in a TDMA frame. We presented the schedulability condition for the case in which messages are transmitted following a FIFO policy. This scheduling discipline is quite simple and requires little processing within the underwater nodes. However, better results may be obtained if some RT priority policies are implemented (as future work). The second solution is based on a heuristic approach. Messages are scheduled following a per-link approach and finding the route with lower delay. This solution improves the 2-step approach of finding the SPF first, for allocating the slots within the frame later. As the heuristic only considers the messages actually being transmitted, unnecessary restrictions are avoided. We also presented a real application (tsunami early alert) in which RT transmissions are necessary.

Acknowledgement. This work has been partially supported by the Spanish MCI and FEDER funds of the EU under contracts TIN2013-44375-R, TIN2013-47245-C2-1-R, TIN2013-47245-C2-2-R, and also the Community Networks Testbed for the Future Internet (CONFINE) project: FP7-288535.

References

[1] R. Santos, J. Orozco, M. Micheletto, S.F. Ochoa, R. Meseguer, P. Millán, C. Molina. Scheduling Real-Time Traffic in Underwater Acoustic Wireless Sensor Networks. *Submitted to UCAmI'16*, Spain, 2016.

[2] R. Santos, J. Orozco, S.F. Ochoa, R. Meseguer, G. Eggly. A MAC Protocol for Underwater Sensors Networks. *Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information*, vol. 9454,2015.

[3] Y. Guan, C.C. Shen, J. Yackoski. Mac scheduling for high throughput underwater acoustic networks. *Wireless Communications and Networking Conference*, 2011.

Making Decisions with Semantic Criteria

Miriam Martínez-García [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
miriam.martinez@urv.cat

Abstract. In some applications, semantic information has an important role in the decision process. This is the case of recommender systems in Tourism [2]. When a tourist has to decide her destination, textual characteristics (i.e. tags) of the different places are key elements to be taken into account (e.g. types of activities to do, main landmarks, etc.). The appropariate analysis of this kind of information is crucial in the development of a new generation of semantic recommender systems. This thesis has the goal to develop new decision aiding tools that incorporate semantic criteria together with numerical ones. Semantic criteria are categorical multi-valued variables whose values are tags (i.e. terms or words) that can be interpreted at a conceptual level. Ontologies are knowledge representation structures that enable this semantic interpretation by exploring the relationships between the terms found in the semantic variable. Furthermore, they may also be used to store the user's preferences. The evaluation of the new semantic knowledge management methods proposed in this thesis will be done in collaboration with the Scientific and Technological Research Park in Tourism and Leisure (Vila-Seca, Tarragona).

1 Introduction

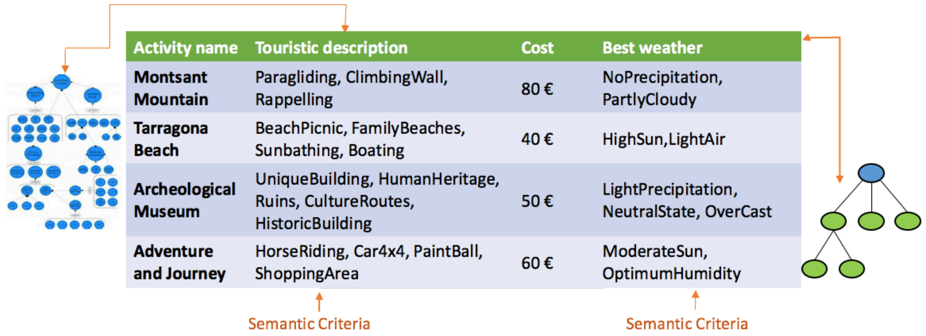
Multi-criteria decision aiding (MCDA) is a well-established discipline focused on proposing decision support tools for the case of dealing with multiple and conflicting criteria [3]. The decision problem may be defined as a choice, ranking or sorting of a set of alternatives, based on their performance on a set of criteria. Although the classic methods were based on operational research and economic theories, nowadays they integrate techniques from other fields, specially from Artificial Intelligence. There are three main methodological approaches in MCDA: utility theory, outranking methods and rule-based systems.

This thesis is focused on the outranking method ELECTRE-III, which is based on the construction of a pairwise outranking matrix that represents the preference structure among a set of alternatives. Once the outranking matrix has been calculated, different exploitation procedures allow to make a

[*] PhD advisors: Aida Valls (URV), Antonio Moreno (URV).

choice (i.e. select the best alternatives), to rank all alternatives or to sort the alternatives into some predefined and ordered classes.

Up to now outranking methods have considered mainly numerical and ordinal scales in the set of criteria. In this PhD thesis I study the use of criteria built upon semantic variables, including additional domain knowledge by means of a background domain ontology. The ontology is a knowledge representation structure that enables an exploration of semantic relationships between the tags found in the semantic variable. We enlarge the ontology base data with information about the user's preferences on the tags. Figure 1 illustrates an example of a recommender system for tourists visiting the province of Tarragona. The alternatives under consideration are a set of activities that can be done in this territory. Each activity is described by a numerical criterion (the cost of the activity) and two multi-valued semantic ones (a description of the categories in which the activity may be classified and a list of the most adequate types of weather to perform the activity). Depending on the interests (i.e. preferences, goals) of the tourist, the best alternative will be a different one. For a family interested on cultural events coming on a rainy day, the archeological museum should be the most suitable option, whereas for an sportive young guy with no money restrictions, some extreme sports on the Montsant mountain could be recommended. Notice that each semantic criterion can use a different domain ontology as support for the analysis of the tags.



Activity name	Touristic description	Cost	Best weather
Montsant Mountain	Paragliding, ClimbingWall, Rappelling	80 €	NoPrecipitation, PartlyCloudy
Tarragona Beach	BeachPicnic, FamilyBeaches, Sunbathing, Boating	40 €	HighSun,LightAir
Archeological Museum	UniqueBuilding, HumanHeritage, Ruins, CultureRoutes, HistoricBuilding	50 €	LightPrecipitation, NeutralState, OverCast
Adventure and Journey	HorseRiding, Car4x4, PaintBall, ShoppingArea	60 €	ModerateSun, OptimumHumidity

Fig. 1: Example of a data matrix with semantic and numerical criteria

2 Including semantic criteria in the ELECTRE method

The first task of the thesis has been the redefinition of the procedure for constructing a valued outranking relation in ELECTRE from semantic multi-valued variables. For each alternative, a semantic variable may have a list of tags (i.e. terms). The concepts of a background ontology correspond to these tags in order to be able to compare them semantically, using some appropriate ontology-based semantic similarity measure.

ELECTRE is a well-known decision aiding method that constructs a valued outranking relation using pseudo-criteria using some discrimination thresholds to manage the uncertainty on the data [3]. For a pair of alternatives, a and b, the procedure for calculating the credibility of the outranking relation aSb is based on two indices inspired on social voting mechanisms: concordance (i.e. overall majority support to aSb) and discordance (i.e. respect to minority opinions against aSb). For numerical and ordinal criteria, the strength of the partial concordance or discordance about aSb for a certain criterion g_j is obtained by comparing the performance of both alternatives $g_j(a)$ and $g_j(b)$. The partial concordance index takes into account two discrimination thresholds, q_j (indifference) and p_j (preference), while the discordance index has a veto v_j threshold to determine the degree of opposition to the assertion aSb. We have proposed a redefinition of the partial concordance and discordance indices to tackle the case of semantic criteria. First, we defined the Tag Interest Score $TIS(c)$ of a concept, which is a numerical value from 0 to 1 that represents the suitability of the concept c for a certain user. Second, the Semantic Win Rate $SWR_j(a, b)$ is defined as a numerical value that indicates the degree of preference of alternative a with respect to b on the semantic criterion g_j. It is based on the evaluation of the user's preferences (from the ontology) about the two sets of tags $g_j(a) = \{t_{1,a}, t_{2,a}, t_{3,a}, ..., t_{|g_j(a)|,a}\}$ and $g_j(b) = \{t_{1,b}, t_2, b, t_{3,b}, ..., t_{|g_j(b)|,b}\}$. As it is a rate that represents the comparison of the performance of a over b, the ELECTRE-III thresholds are now defined as follows:

- μ_j is the minimum value for the strength of $SWR_j(a, b)$ to consider a maximum concordance with aSb.
- p_j indicates the maximum difference between $SWR_j(a, b)$ and μ_j that still shows some preference of a with regards to b, thus still supporting the relation aSb to a certain degree.
- v_j is the veto threshold, which shows the minimum negative difference between $SWR_j(a, b)$ and μ_j that requires the full discordance with the outranking relation.

3 Constructing a semantic user profile

User profiling is required in many decision support systems. Nowadays it is becoming more common to find decision problems involving non-numerical data, such as multi-valued *semantic criteria*, which take as values the concepts of a given domain ontology. Different models of representation of the preferences have been revised [1]. Upon that, we propose to create a *semantic user profile* by storing preference scores into the ontology. This preferential information can be later exploited to rank and recommend the most suitable alternatives for each user with the ELECTRE method (or other decision aiding methods).

As said before, ontologies store domain information in the form of concepts and taxonomic relations. In this work we propose to include a numerical interest score attached to the most specific concepts (i.e. the leaves of the taxonomy). With this score, which is associated to very detailed concepts, we are able to distinguish better the preferences of the user, improving the quality of the decision. Provided that ontologies usually have hundreds of concepts, it is not feasible to obtain all their scores at the beginning. Therefore, given a concept c with an unknown preference, an inference procedure has been designed to estimate it. The basic idea is to find a subset of concepts semantically similar to c and to aggregate their scores. After studying the literature on aggregation operators [6], we propose using the *WOWA (Weighted Ordered Weighted Average)* operator with two weighting factors: *OWA* weights define the pessimistic/optimistic aggregation policy, while criteria weights give different importance to the aggregated values in terms of their semantic distance to c [5]. We are currently studying which are the most appropriate parameters of WOWA depending on the structure of the ontology and the number of missing tag interest scores.

Acknowledgement. This PhD thesis is being developed with the support of a research grant Martí Franquès (2014PMF-PIPF-70) and the research project SHADE (TIN-2012-34369)

References

[1] A. Valls, A. Moreno, J. Borràs. "Preference representation with ontologies". *Multicriteria Decision Aid and Artificial Intelligence.*, pp 77-99, 2013.

[2] J. Borrás, A. Moreno, A. Valls. "Intelligent tourism recommender systems: a survey". *Expert Systems with Applications*, vol. 41(16) pp. 7370-7389, 2014.

[3] J.Figueira, S. Greco, M.Ethrogott. "Multiple Criteria Decision Analysis: State of the Art Surveys ". *International Series in Operations Research & Management Science*, vol.78. Springer, 2005.

[4] M. Martínez-García, A. Valls, A. Moreno. "Construction of an outranking relation based on semantic criteria with ELECTRE-III ". *Information Processing and Management of Uncertainty in Knowledge Based Systems*, In press, 2016.

[5] M. Martínez-García, A. Valls, A. Moreno. "Using aggregation operators to infer semantic preferences ". *28th European Conference on Operational Research.*, In press, 2016.

[6] V. Torra. "The WOWA operator: a review ". *Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice.*,vol. 254 pp. 17-28. Springer, 2011.

Efficient Wireless Channel Characterization for Context-Aware Scenarios

Fran Casino

Smart Health Research Group. Department of Computer Engineering and Mathematics, Rovira i Virgili University.
Av. Països Catalans 26. 43007 Tarragona. Catalonia. Spain
Corresponding author e-mail: franciscojose.casino@urv.cat

1 Introduction

Nowadays, the population shift from rural to urban areas poses severe challenges to cities. In big urban areas, factors related to economies of scale help to reduce operational costs. However, managing such cities is challenging due to the large number of inhabitants and their needs. Thus, such management procedures have to be adapted to a growing and very demanding population. The citizens' quality of life is one of the most relevant aspects in this scenario. Therefore, healthcare services are also taken into account, due to their relevance and the potentially high cost inherent in service provision, which is constantly increasing due to the population growth and the increase in life expectancy. Smart health (s-health), which is understood as a natural evolution of e-health in the context of smart cities, was introduced by Solanas et al., and can be defined as follows:

> *"Smart health (s-health) is the provision of health services by using the context-aware network and sensing infrastructure of smart cities."*
> Solanas et al. [13]

However, such new health paradigm has evolved to transcend the boundaries of smart cities and to be fully applicable to any context-aware environment with the aim to improve the quality of life of people. Noticeably, s-health is a subclass of e-health because it is founded in ICT like mobile health (m-health). However, it differs from m-health in that the underlying infrastructure is not necessary mobile and in most cases it is static. With the aim to provide a long-term sustainable healthcare system, optimized organization and management systems are combined in novel healthcare service implementations. In this way, new models based on e-health, m-health, s-health and Ambient Assisted Living, in which mobile communication systems, distributed sensor networks and optimized middleware have been proposed and are currently being deployed following different strategies (*i.e.* in conjunction with telecom operators, hybrid mobile WLAN-WSN implementation, etc.). Therefore, one of the main enablers of context aware environments is the use of different wireless communication systems, which provide seamless connectivity to a potentially very large set of transceivers embedded in mobile and wearable terminals, WLAN hot spots or dense wireless sensor network deployments.

2 Wireless Channel Characterization

In the last decades, the use of wireless systems has increased substantially, given the popularity of mobile networks, wireless LAN and wireless sensor networks. The advent of context-aware environments, mainly driven by the trend in smart city/smart region development, is going to increase furthermore the deployment of 4G mobile networks, IoT and the overall evolution towards high capacity and capillarity of 5G systems. In this scenario, one of the main considerations is to control interference precisely, in order to increase coverage/capacity ratios. In this sense, given the wide variety of existent wireless systems and the inherent complexity of large, dense urban scenarios, radio-planning tasks are compulsory to fully account for useful server signals as well as intra-system and inter-system interference sources. In order to accomplish that, several techniques can be used, from semi-empirical regressive methods, which exhibit large errors and measurement dependent models, to deterministic based techniques like full wave electromagnetic simulation. As a midpoint between precision and computational cost, deterministic Ray Launching (RL) methods offer a good trade-off between precision and computational cost. However, performing real measurements is very time consuming and becomes impractical in complex, large scenarios [1]. With the aim to avoid this burden, simulation techniques based on Ray Tracing, combining Geometric Optics and Uniform Theory of Diffraction, are used to predict waves' behaviour within a given environment. Those simulations depend on a number of parameters, namely angular resolution, number of rebounds, cells size, etc. By tuning these parameters, high-definition (HD) and low-definition (LD) results can be obtained. Although more practical than manual measurements, the computational cost of simulations in HD prevents their use in complex environments and their LD counterparts are applied.

3 Recommender Systems and Collaborative Filtering

Recommender systems [11] play an active role in the Internet through the advances in data mining and artificial intelligence. Collaborative Filtering (CF) [8] is a kind of recommender system that comprises a large family of recommendation methods. The aim of CF is to make suggestions on a set of items I (*e.g.* restaurants, films or routes) based on the preferences of a set of users U that have already acquired and/or rated some of those items. Recommendations provided by CF methods are based on the premise that similar users are interested in similar items (*i.e* they share similar patterns). Therefore, items which pleased user u_a could be recommended to user u_b, if u_a and u_b are similar. In order to predict whether an item would interest a given user, CF methods rely on a matrix M of n users (rows) and m items (columns), where each matrix cell $M_{i,j}$ stores the rate of user i on item j. The interested reader could refer to [5,6,7,12] for a detailed CF's state-of-the-art.

4 Hybrid method for Wireless Channel Characterization

One of the issues to consider in the design of communication networks in the context of s-health scenarios is their performance in terms of coverage/capacity ratios [4,9], with particular consideration of the impact of interference due to simultaneous use of multiple users and systems [10]. It is in this case where careful radiofrequency signal analysis, in terms of useful signal transmission and existence of potential interference levels must be estimated, as a function of user density, transceiver type and location. As previously stated, wireless signal analysis in large complex scenarios is computationally costly and requires the use of optimized deterministic techniques. In the case of very large scenarios, such as cities, this approach can still be computationally too demanding and combination with other estimation approaches is compulsory [1]. In order to minimize computational cost for certain scenarios, we proposed the combination of in-house developed 3D Ray Launching code with CF techniques in [2]. The main idea of our hybrid proposal is to use the ability of CF methods of predicting rates to find the values of empty cells in LD simulations. Therefore, we implemented a hybrid method which follows a two step procedure (*i.e.* neighbourhood search and recommendation/prediction computation) in order to estimate the power level of empty/error cells so that they are as similar as possible to the values that would have been obtained in an HD simulation. Figure 1 shows an example of the outcomes obtained by the hybrid approach. Our proposal has been tested in different scenarios such as medical emergency rooms [3], in houses with concrete rooms distribution [10] and at universities with different kinds of laboratories and offices [2]. The outcomes of such experiments showed that our proposal was not only more accurate than other state-of-the-art methods, but also more efficient in terms of computational cost. Therefore, we may conclude that our proposal is accurate as well as efficient, and that could be used to enhance the accuracy of LD simulations in diverse scenarios.

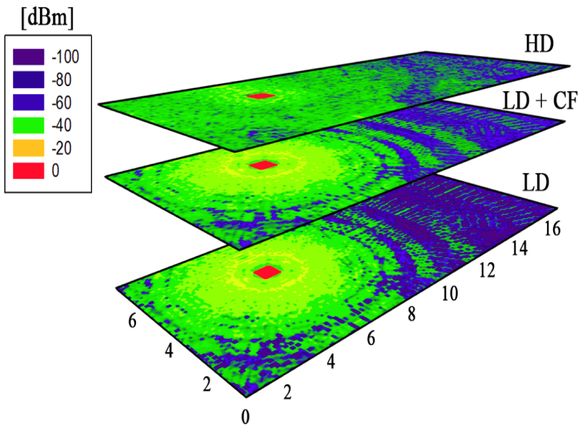


Fig. 1: Received power level estimation when using HD (top), LD+CF (middle) and LD (bottom).

5 Conclusions

In this paper, we have stated the importance of healthcare systems in the context of smart cities. Moreover, we have pointed out the relevance of the new health paradigm, smart-health, and the importance of increasing the coverage/capacity ratios of wireless systems in context-aware scenarios such as dense urban areas or hospitals. We have showed that CF methods can be successfully applied to improve the accuracy of LD simulations performed by a RL approach and that our proposal is fast and could help to reduce the HD simulation costs. Therefore, Collaborative Filtering could be integrated with the sensing infrastructure of smart cities to improve the sustainability by optimizing the resource usage in the communication networks field.

Acknowledgments and disclaimer

This work is partly supported by La Caixa Foundation through Project SIMPATIC and by the Martí-Franquès Research Grants Programme. We would also like to thank our colleagues from the Electrical and Electronic Engineering Department of the Public University of Navarra who provided insight and expertise that greatly assisted the research presented in this paper.

References

[1] L. Azpilicueta, M. Rawat, K. Rawat, F. Ghannouchi and F. Falcone, Convergence analysis in deterministic 3D ray launching radio channel estimation in complex

environments, *Applied Computational Electromagnetic Society Journal, Vol. 29, No. 4, pp. 256-271, April (2014)*

[2] F. Casino, L. Azpilicueta, P. Lopez-Iturri, E. Aguirre, F. Falcone, and A. Solanas, Optimized wireless channel characterization in large complex environments by hybrid ray launching collaborative filtering approach, *Tech. Rep.*, February 2015. [Online]. Available: http://s-health.eu/publications/technical_report_optimized_ray_launching.pdf

[3] F. Casino, L. Azpilicueta, P. Lopez-Iturri, E. Aguirre, F. Falcone, and A. Solanas, Hybrid-based optimization of wireless channel characterization for health services in medical complex environments, *6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2015.

[4] F. Casino, P. Lopez-Iturri, E. Aguirre, L. Azpilicueta, A. Solanas and F. Falcone, Dense wireless sensor network design for the implementation of smart health environments, *International Conference on Electromagnetics in Advanced Applications (ICEAA)*, pp. 752-754, 2015

[5] F. Casino, C. Patsakis, D. Puig, and A. Solanas, On privacy preserving collaborative filtering: Current trends, open problems, and new issues, *ICEBE*, pp. 244-249. (2013).

[6] F. Casino, J. Domingo-Ferrer, C. Patsakis, D. Puig, and A. Solanas, Privacy preserving collaborative filtering with k-anonymity through microaggregation, *ICEBE*, pp. 490-497. (2013).

[7] F. Casino, J. Domingo-Ferrer, C. Patsakis, D. Puig, and A. Solanas, A k-anonymous approach to privacy preserving collaborative filtering, *Journal of Computer and System Sciences, Vol. 81, Issue. 6, pp. 1000-1011*, Dec. 2014.

[8] D. Goldberg, D. Nichols, B. M. Oki, D. Terry, Using collaborative filtering to weave an information tapestry, *Communications of the ACM*, 35(12), 61-70. (1992).

[9] A. de Lejarazu, P. Lopez-Iturri, E. Aguirre, L. Azpilicueta, F. Falcone, F. Casino and A. Solanas, Challenges in the implementation of context-aware scenarios within emergency rooms, *6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2015.

[10] P. Lopez-Iturri, F. Casino, E. Aguirre, L. Azpilicueta, F. Falcone, and A. Solanas, Performance analysis of zigbee wireless networks for aal through hybrid ray launching and collaborative filtering, *Journal of Sensors*, vol. 2016, no. 2424101, pp. 1-16, 2016.

[11] P. Resnick, H. Varian, Recommender systems, *Communications of the ACM*, 40(3), 56-58. (1997).

[12] Y. Shi, M. Larson, A. Hanjalic, Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges, *ACM Computing Surveys (CSUR)*, 47(1), 1-45. (2014).

[13] A. Solanas, C. Patsakis, M. Conti, I. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. Pérez-Martínez, R. Di Pietro, D. Perrea, and A. Martínez-Ballesté, Smart health: A context-aware health paradigm within smart cities. *IEEE Communications Magazine, Vol. 52, No. 8, pp. 74-81*, August. (2014).

Reducing Network Overhead on Personal Cloud Systems

Raúl Sáiz-Laudó [⋆]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
raul.saiz@estudiants.urv.cat

1 Introduction

As a tool for personal storage, file synchronization and data sharing, cloud storage services such as Dropbox, Box, Google Drive, etc., have quickly gained popularity in the last few years. This services provide users with reliable data storage that can be automatically synced across multiple devices, and also shared among a group of users. The main problem with these services is that they rely on the client-server communication paradigm to make their content available. And depending upon the scenario, it may introduce a huge network overhead.

To minimize network overhead, cloud storage services use a variety of techniques such as binary diffs on file chunks, file bundling, data compression, among others. One of those techniques is *sync deferment*, which consists in aggregating several file mutations in a single message to improve network communication. In practice, it has been implemented using a *fixed* time threshold of T seconds (Dropbox) [1]: *once T have elapsed, the client triggers an update to the server.* The disadvantage of *fixed* sync deferment is that it is limited in terms of usage scenarios. For instance, consider a collaborative document editing scenario where the frequency of modifications to a file are huge. Just in Dropbox, for 8.5% of its users, sync traffic caused by frequent modifications represents more than 10% of their total network traffic [2]. In settings like this, a "smarter" sync deferment mechanism is actually necessary to decrease the overhead generated by the amount of superfluous data sent by clients over time.

In particular, the authors of the recent measurement on cloud storage services [1] suggest the usage of *adaptive* sync deferment techniques to overcome the limitations found in static techniques.

[⋆] Marc Sánchez-Artigas , Pedro García-López

2 Adaptive Sync Deferment (ADS)

The adaptive sync deferment algorithm introduced in [1] proposes to adaptively tune the sync deferment time T_i to follow the latest file update. Simply put, when updates happen more frequently, the idea is that T_i becomes shorter to keep pace with a higher update rate. And when they happen less frequently, it gets longer. To achieve this, T_i is adapted in the following simple iterative manner:

$$T_i = \min\left(\frac{T_{i-1}}{2} + \frac{\Delta t_i}{2} + \epsilon, T_{max}\right), \tag{1}$$

where Δt_i is the inter-update time between the $(i-1)$-th and the i-th data updates, and $\epsilon \in (0,1.0)$ is a small constant that guarantees T_i to be slightly longer than Δt_i in a small number of iteration rounds. T_{max} is also a constant representing the upper bound on T_i. Note that a too long T_i will harm user experience by introducing unacceptable long sync delays.

3 Our Proposal: Rate-based Sync Deferment (RDS)

Although the authors of [1] demonstrate that Google Drive and OneDrive will receive a negligible overhead close to 1 by using the ADS algorithm, this algorithm does not take into account the amount of bytes triggered by each update. Data volume is also an important factor. To better understand this, consider a regular file update pattern. While ADS will be able to easily adapt to the sync deferment time T_i to this pattern, it would not be able to keep the overhead low if the data volume per update was very variable, for instance, by sinusoidally oscillating over time.

For this reason, we have added the concept of *Rate* in our algorithm, measured in bytes per second, to do the calculations that determine whether a batch of updates should be pushed or not to the cloud. Observe that the introduction of the notion of *Rate* provides us with a finer-grained control of the network overhead, for we have called our algorithm RDS.

More specifically, the $Rate_i$ is calculated from the amount of bytes since the last update ΔB_i and the inter-update time Δt_i. Then,

$$Rate_i = \frac{\Delta B_i}{\Delta t_i}.$$

Note that ΔB_i is easy to calculate by comparing the file chunks at each update time t_i. Equipped with this information, T_i is adapted in an iterative manner as follows:

$$T_i = \min(T_{Rate_i}, T_{max}), \tag{2}$$

where T_{max} is a constant representing the upper bound of T_i (i.e., upper bound on the unsynced time) and T_{Rate_i} is computed as follows:

$$T_{Rate_i} = TR_{i-1} * \alpha + \Delta TR_i * (\alpha - 1) + \epsilon,$$

where $TR_i = \overline{B}/Rate_i$, $\overline{B}$ is the maximum number of bytes required to meet our targeted overhead objective, ϵ is a small constant and α is a weighting factor (EWMA). More specifically, α determines the agility of the time estimator T_{Rate_i} in following the abrupt changes in the actual data rate or the stability of the estimator in ignoring short term variations.

4 Preliminary results

To evaluate the effectiveness of RDS, we have conducted several experiments on both real-life and synthetic workloads. To provide a broader view, we have compared the efficacy of RDS with the ADS algorithm.

Workloads. We have considered the following workloads:

- **UB1.** We have used a real log file of user file synchronization mechanism during a whole day from *Ubuntu One* (UB1) platform [3]. Only users with relevant activity periods were chosen for the tests. By "activity period" we mean the time elapsed between the first data update in the day and the last one. Users displaying activity periods of less than 1 hour were considered to be irrelevant for our study. Further, activity periods were split into *sessions*. We considered as a session a sequence of file updates where the inter-update time Δt_i was lower than 900 seconds. When this condition was not satisfied, the current session was considered to be "close", so the next data update received became the first one in the new session.

  We have extracted two sessions: A regular pattern with a coefficient of variation (CV) in the data volume per update lower than 1 (Session A), and a pattern with high variability, i.e., with the $CV > 1$ (Session B).
- **Synthetic**. As another pattern, we have artificially generated a triangular pattern that triggers a file mutation every 5 seconds. More specifically, an initial write of 10KBs is performed, and subsequently, 10KBs of new data is appended at every new update until the file size reaches a size of 80 KBs. From that point onwards, a decremental update of 10 KBs is done until the file size becomes 10KBs again.

  This is repeated over time, mimicking a triangular signal.

Results. Table 1 reports the results for all the three workloads. The metrics for the comparison were: the resulting network overhead, the number of uploads to the cloud servers during a session, and the average sync deferment time obtained in practice.

As shown in Table 1, ADS does not work well in all the cases, as it is possible to find real and synthetic workloads where the sync deferment time becomes extremely long, impacting user experience very negatively and

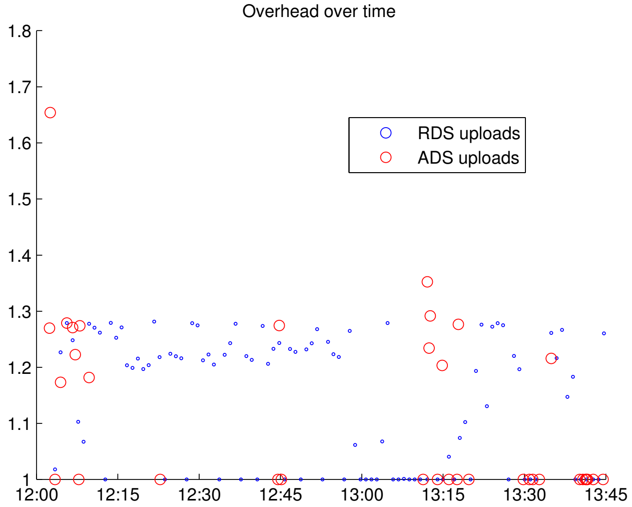


Fig. 1: Session B. Evolution of the network overhead over time.

Table 1: Comparison between ADS and RDS.

	Algorithm	Overhead	# Uploads	Sync Deferment Time
Session A	RDS	1.0137	123	33.447 seconds
	ADS	1.0456	164	40.650 seconds
Session B	RDS	1.1416	99	13.120 seconds
	ADS	1.1159	36	351.34 seconds
Synthetic	RDS	1.2038	4664	7.9991 seconds
	ADS	1.3513	3	∞

increasing the frequency of user conflicts when they concurrently edit a file. In contrast, RDS delivers an equivalent network overhead in all the scenarios, but it is able to keep a good synchronization level at the same time, making RDS more stable and responsive. To better understand this, the instantaneous overhead for every update is shown in Fig. 1 for the Session B. As can be shown in this figure, RDS waits less before triggering an update to the cloud, exhibiting a smoother behavior.

References

[1] Zhenhua Li, Cheng Jin, Tianyin Xu, Christo Wilson, Yao Liu, Linsong Cheng, Yunhao Liu, Yafei Dai, and Zhi-Li Zhang. Towards Network-level Efficiency for Cloud Storage Services. In Proceedings of the 2014 Conference on Internet Measurement Conference (IMC '14). ACM, New York, NY, USA, 115-128 , 2014.

[2] Z. Li, C. Wilson, Z. Jiang, Y. Liu, B.Y. Zhao, C. Jin, Z.-L. Zhang, and Y. Dai. Efficient Batched Synchronization in Dropbox-like Cloud Storage Services. In Proc. of Middleware, pages 307–327. Springer, 2013.

[3] Raúl Gracia-Tinedo, Yongchao Tian, Josep Sampé, Hamza Harkous, John Lenton, Pedro García-López, Marc Sánchez-Artigas, and Marko Vukolic. Dissecting UbuntuOne: Autopsy of a Global-scale Personal Cloud Back-end. In Proceedings of the 2015 ACM Conference on Internet Measurement Conference (IMC '15). ACM, New York, NY, USA, 155-168. 2015.

Privacy-Preserving Statistical Computation in the Cloud

Sara Ricci [⋆]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
sara.ricci@urv.cat

Abstract. We tackle the problem of privacy-preserving statistical computation in the cloud. The goal is to use the cloud not only to store sensitive data but also to perform computations on them. Specifically, we focus on protocols to obtain the sample covariance matrix of the sensitive numerical data set, and on protocols to obtain the contingency matrix and the distance covariance matrix of the sensitive categorical data set, calculations that underlie most statistical analyses. However, the multi-cloud is semi-honest, that is, it follows the protocols but is not authorized to learn the sensitive data. We rely on the use of several clouds; if these can be assumed not to collude, we use vertical data splitting among clouds; if clouds may collude, we present two alternative protocols that withstand collusion at the expense of increased cloud storage.

1 Introduction

Data have become a crucial asset of many enterprises, organizations and public administrations. Collecting and analyzing large amounts of data related to individuals does not only improve research, but it also drives a tremendous business [13]. However, local storage and processing of such big data is often unfeasible for the data controllers because of the associated costs (software, hardware, energy, maintenance). An attractive possibility for a data controller is to outsource data to a cloud [2]. This brings several benefits such as large and highly scalable storage/computation resources at a low cost and with ubiquitous access. On the other hand, concerns about security and privacy still have a detrimental impact on the adoption and acceptance of cloud services: neither users, nor companies want the cloud service provider (CSP) to read, use or sell their data.

In this context, the need emerges to find a secure, efficient and privacy-preserving storage and processing methods for the (sensitive) data outsourced to the cloud. This is precisely the main goal of the European project CLARUS [5] which consists in a proxy located in a domain trusted by the

[⋆] PhD advisor: Josep Domingo Ferrer

data controller (e.g., a server in her company's intranet or a plug-in in her device) that implements security and privacy-enabling features towards the CSP so that i) the CSP only receives privacy-protected versions of the controller's data, ii) CLARUS makes the access to such data transparent to the controller's users (by adapting their queries and reconstructing the results retrieved from the cloud) and iii) it remains possible for the users to leverage the cloud to perform accurate computations on the outsourced data without downloading them.

To do so, CLARUS particularly relies on *data splitting* as a data protection technique: data are partitioned into several fragments, each of which is stored in the *clear* in a cloud provided by a different CSP [4]. Data splitting is an alternative that is more efficient and functionality-preserving than encryption-based methods (e.g., CipherCloud, PerspecSys, SecureCloud, etc.). In general, even though searchable and homomorphic encryption allow performing some operations on ciphertext [7], computing on encrypted data is extremely limited and costly [9], and it requires careful management of encryption keys. In contrast, the *vertical* data splitting implemented by CLARUS protects privacy (confidential information on an individual is partitioned into fragments that cannot be linked) and allows computation to be performed on clear data.

2 Obtained Results

In [3], we evaluated several non-cryptographic proposals for statistical computation (basically correlations) on split data, and we enhanced and proposed some protocols adapted to the CLARUS scenario. In [10], we extended these results by considering also cryptographic protocols and by relaxing the non-collusion assumption. We first assumed that the CSPs do not collude to reconstruct the original data from the fragments, and we presented two protocols for this setting. We then relaxed the non-collusion assumption and present two protocols that are collusion-resistant, even though they require substantial cloud storage (because they rely on data replication rather than splitting). In both the articles, we focused on the computation of the sample covariance matrix because many statistical analyses, such as regression, classification, principal component analysis, etc., are based on it.

All these protocols and methods were designed for numerical data. However, many of the (personal) data currently collected from a variety of sources (social networks, surveys, B2C transactions, etc.) are not numerical. In [11], we adapted some of the methods proposed for split numerical data to categorical data. Specifically, we described two protocols (with and without cryptography, respectively) to compute the contingency table (used for the χ^2-test of independence [1]) and the distance covariance matrix (based on the more recent *distance covariance/correlation* measure, see [12]) needed to measure the correlation between two categorical attributes stored in different clouds.

In all the articles, we compared the computational and communication costs of the described protocols against a benchmark consisting of the CLARUS proxy downloading the entire data set and locally computing on the downloaded data set. If clouds can be assumed not to collude, data splitting is probably the best choice, due to simplicity and flexibility.

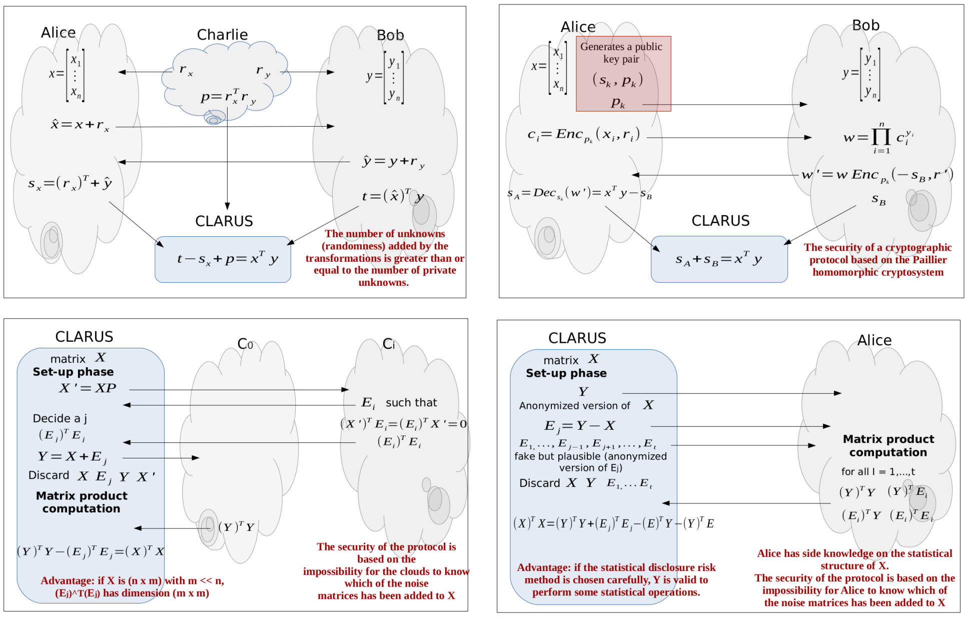

Alice
Charlie
Bob
$x=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}$
r_x
r_y
$p=r_x^T r_y$
$y=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}$
$\hat{x}=x+r_x$
$\hat{y}=y+r_y$
$s_x=(r_x)^T+\hat{y}$
$t=(\hat{x})^T y$
CLARUS
$t-s_x+p=x^T y$
The number of unknowns (randomness) added by the transformations is greater than or equal to the number of private unknowns.
Alice
Bob
$x=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}$
Generates a public key pair
(s_k, p_k)
p_k
$y=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}$
$c_i=Enc_{p_k}(x_i,r_i)$
$w=\prod_{i=1}^{n} c_i^{y_i}$
$w'=w\,Enc_{p_k}(-s_B,r')$
$s_A=Dec_{s_k}(w')=x^T y-s_B$
s_B
CLARUS
$s_A+s_B=x^T y$
The security of a cryptographic protocol based on the Paillier homomorphic cryptosystem
CLARUS
C0
Ci
matrix X
Set-up phase
$X'=XP$
E_i such that
$(X')^T E_i=(E_i)^T X'=0$
Decide a j
$(E_j)^T E_i$
$(E_i)^T E_i$
$Y=X+E_j$
Discard X E_j Y X'
Matrix product computation
$(Y)^T Y$
$(Y)^T Y-(E_j)^T E_j=(X)^T X$
Advantage: if X is (n x m) with m << n, (Ej)^T(Ej) has dimension (m x m)
The security of the protocol is based on the impossibility for the clouds to know which of the noise matrices has been added to X
CLARUS
Alice
matrix X
Set-up phase
Y
Anonymized version of X
$E_j=Y-X$
$E_{1},\dots,E_{j-1},E_{j+1},\dots,E_t$
fake but plausible (anonymized version of Ej)
Discard X Y $E_{1},\dots E_t$
Matrix product computation
for all I = 1,...,t
$(Y)^T Y$ $(Y)^T E_i$
$(E_i)^T Y$ $(E_i)^T E_i$
$(X)^T X=(Y)^T Y+(E_j)^T E_j-(E)^T Y-(Y)^T E$
Advantage: if the statistical disclosure risk method is chosen carefully, Y is valid to perform some statistical operations.
Alice has side knowledge on the statistical structure of X.
The security of the protocol is based on the impossibility for Alice to know which of the noise matrices has been added to X

Fig. 1: Protocols for the secure scalar product. The top charts work in the non-colluding scenario and bottom charts in the collusion-resistant scenario.

3 Computation on vertically partitioned data

In vertical splitting, analyses that involve only attributes in a single fragment are really fast and easy to compute: the cloud storing the fragment can compute and send the output of the analysis to the CLARUS proxy. Unfortunately, sample covariance matrix, the contingency matrix and the distance covariance matrix involve attributes stored in different fragments, and thus communication between clouds. Obtaining the sample covariance matrix (the contingency matrix and the distance covariance matrix, respectively) in vertical splitting among several clouds can be decomposed into several secure scalar products to be conducted between pairs of clouds (see [10] and [11]). Secure scalar products can be based on cryptography (the protocol in [8] involves homomorphic encryption), or not ([6], modify the data before sharing them in such a way that the original data cannot be deduced from the shared data but the final results are preserved). See Figure 1 for a sketch of the secure scalar product protocols.

Acknowledgement. The author is supported by an FPI grant from the Spanish Government. Partial support to this work has been received from the European Commission (projects H2020-644024 "CLARUS" and H2020-700540 "CANVAS"), from the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2014 SGR 537), and from the Spanish Government (project TIN2014-57364-C2-1-R "SmartGlacis" and TIN 2015-70054-REDC). The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

[1] Agresti, A., Kateri, M.: Categorical Data Analysis. Springer (2011).

[2] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: A view of cloud computing. Communications of the ACM, vol. 53, no. 4, pp. 50–58 (2010).

[3] Calviño, A., Ricci, S., Domingo-Ferrer, J.: Privacy-preserving distributed statistical computation to a semi-honest multi-cloud. In IEEE Conf. on Communications and Network Security (CNS 2015). IEEE (2015).

[4] Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Selective data outsourcing for enforcing privacy. Journal of Computer Security, vol. 19, no. 3, pp. 531–566 (2011).

[5] CLARUS - A Framework for User Centred Privacy and Security in the Cloud, H2020 project (2015-2017). `http://www.clarussecure.eu`

[6] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., Zhu, M.: Tools for privacy preserving distributed data mining. ACM SiGKDD Explorations Newsletter, vol. 4, no. 2, pp. 28–34 (2002).

[7] Dubovitskaya, A., Urovi, V., Vasirani, M., Aberer, K., Schumacher, M.: A Cloud-Based eHealth Architecture for Privacy Preserving Data Integration. In ICT Systems Security and Privacy Protection. Springer, pp. 585–598 (2015).

[8] Goethals, B., Laur, S. , Lipmaa, H., Mielikäinen, T.: On private scalar product computation for privacy-preserving data mining. In Information Security and Cryptology - ICISC 2004, Lecture Notes in Computer Science, vol. 3506, pp. 104–120, Springer (2005).

[9] Ren, K., Wang, C., Wang, Q.: Security challenges for the public cloud. IEEE Internet Computing, no. 1, pp. 69–73 (2012).

[10] Ricci, S., Domingo-Ferrer, J., Calviño, A., Domingo-Enrich, C.: Privacy-Preserving Statistical Computation in the Cloud. Manuscript.

[11] Ricci, S., Domingo-Ferrer, J., Sánchez, D.: Privacy-Preserving Cloud-Based Statistical Analyses on Sensitive Categorical Data. In press in MDAI2016, Springer LNAI series.

[12] Székely, G.J., Rizzo, M.L.: Brownian distance covariance. The Annals of Applied Statistics 3.4: 1236-1265 (2009).

[13] U.S. Federal Trade Commission: Data Brokers, A Call for Transparency and Accountability (2014).

A Numerical Study on Bayesian Search Strategies

Lluís Arola Fernández [⋆] [⋆⋆]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
lluisarolaf@gmail.com

1 Introduction

The search of a lost target in a spatial domain is a physical process that appears in many different contexts [1,3]. Search Theory is a broad field in applied mathematics that delves into the study of search strategies, either from the analytical derivation of optimal trajectories [6] or the study of performance and behavior of the strategies existing in nature [1,2,3].

The current work explores the search problem from a numerical approach. The construction of a Bayesian Search Algorithm with an Information Gain Maximization Criteria defines an adequate framework to analyze the trajectory and performance of several strategies and to identify the different deviations from optimal plans [2]. We relate the information-processing mechanism to the bounded rationality of the agent [5]: the searcher, as in real situations, shows cognitive biases when dealing with incomplete information problems.

2 The Bayesian Model

The search problem is treated as an informative issue. The uncertainity of the system is captured in a probability distribution function $P(x, y)$ for the target allocation in a given domain. The agent conducts the search by constantly updating the probabilistic map after unsuccessful local searches, applying a Bayesian inference as new evidences are acquired [1]:

$$P(x_j, y_j) = \prod_{i=0}^{j-1} P_{ND}(x_i, y_i) P(x_0, y_0) = \prod_{i=0}^{j-1} [1 - P_D(x_i, y_i)] P(x_0, y_0) \tag{1}$$

where $P(x_j, y_j)$ measures the current probability after j iterations, and $P_D(x_i, y_i)$ is the detection probability which exponentially decreases with the distance between the target and the agent's current position.

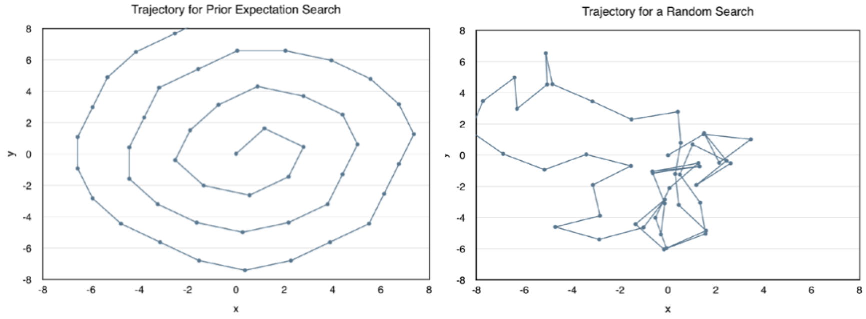
[⋆] PhD advisors: Dr. Sergio Gómez and Dr. Alex Arenas.

[⋆⋆] Dr. Daniel Campos (UAB).

A movement criteria is required to choose the next search regions. We impose the optimal search policy derived by E.T Jaynes in the context of Information Theory [4]. It states that the searcher should maximize the information gain about the target, being able to exploit the prior knowledge as fast as possible to obtain the maximum saving in search effort.

$$\varDelta S = S_0 - S_1 = -\sum^{cells} P_0 ln(P_0) + \sum^{cells} P_1 ln(P_1) \tag{2}$$

In the algorithm, the agent samples S new positions, and it moves towards the one that maximizes the difference of Shannon's Entropy $\varDelta S$ between the current and the future state of the search [1,4].


Trajectory for Prior Expectation Search

Trajectory for a Random Search

Fig. 1: Difference between information gain maximization and random strategies.

At the end, the model attempts to describe the real cognitive process during the search. The agent searches in the most probable region, modifies the prior beliefs and *thinks* where to move next until the target is found [3].

2.1 Parametrization and computational costs

In this model, all the parameters involved in the search can be quantified and classified into dimensional (scale of the problem), strategical (jump width and sampling constraints) and computational (number of cells, samples and iterations). The computational costs of the algorithm can be related to the search time in a real process. When the total time Q is fixed, a trade-off appears [2]: the information-processing mechanism requires some effort to sample the S new positions in detriment of the number of iterations J. We found an analytical expression:

$$Q(N, S, J) \sim \sum_{t=0}^{t=J} SN^2 t = SN^2 \frac{J(J+1)}{2} \tag{3}$$

that relates the effort Q to the computational parameters and shows that an agent's strategy requires an specific effort allocation to conduct the search.

3 The Numerical Study

3.1 State of Knowledge and Effort Allocation

We attempt to reproduce the behavior of a real searcher that is expected to show biases from optimal plans. The deviations might come from the State of Knowledge K (choice of strategical parameters and movement criteria) and the Effort Allocation D (amount of information-processing in terms of sampling).

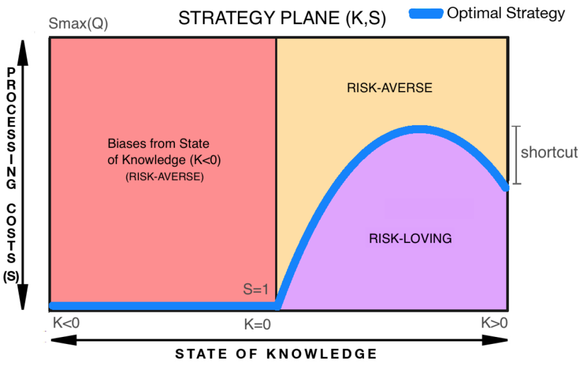


Fig. 2: Deviations from optimal strategies in the plane (K,S).

where $K = 0$ is a reference point for the strategies with no available information (Random Search), and $K \neq 0$ for the strategies that use some (*valuable* or not) information. The numerical study seeks to identify the sources of error coming from bad estimations on the search parameters ($K < 0$), or an inadequate allocation of the effort (risk-averse and risk-loving behaviors).

3.2 Trajectory and Performance Analysis

We analyze numerically the impact of several search parameters under different initial distributions and sampling strategies. The parametric analysis allows us to find the optimal search values and detect the different sources of error in the path-planning [2]. The results show that an appropriate parametrization can increase the performance of the strategies guided by a prior expectation when some useful information of the system is available. However, incorrect estimations on the initial distribution or search parameters can lead to a very bad performance, even surpassed by pure random strategies, which do not follow any movement criteria or sampling method.

3.3 A Search Game for Four Players

We build up four players with different States of Knowledge (from the Pro Player with optimal parameter values to a completely Random Player) in order to understand better the cognitive biases in a search strategy.

Three scenarios are defined, where a large number of targets are hidden in consecutive searches with a fixed time or effort Q, and the performance of the players is measured for different effort allocations (i.e number of samples S).

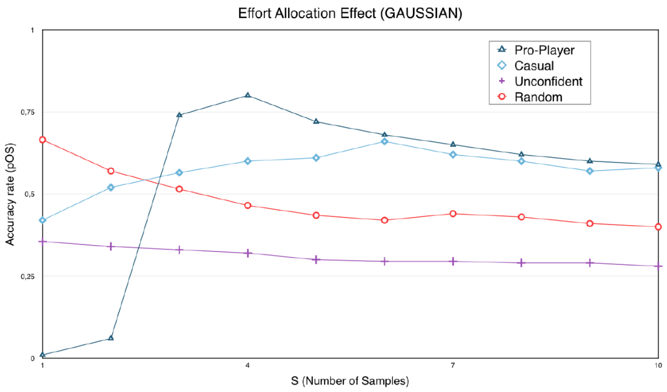


Fig. 3: Players performance for different effort allocations in the Gaussian scenario.

The bayesian model applied in these games validates the hypothesis of the work and shows that each strategy has an specific optimal Effort Allocation Point, which might change depending on the State of Knowledge of the agent and the uncertainty of the system. We have empirically shown the trade-off between information-processing and randomness in a search strategy, where the rationality of the agent is bounded by its theoretical knowledge on the problem and the global effort constraints [2,5].

References

[1] C. Barbieri, S. Cocco, R. Monasson. On the trajectories and performance of Infotaxis. *Physics Bio.* EPL, 94 - 20005, 2011.

[2] D. Campos, V. Mendez, J. Palmer and F. Bartumeus. Path planning in the light of random search theory: coping with human errors and uncertainty. *Not published yet.* 2015.

[3] A. Calhoun, S. Chalasani, T. Sharpee. Maximally informative foraging by Caenorhabditis elegans. *eLIFE 3.* e04220, 2014.

[4] E.T. Jaynes. Entropy and Search Theory. *First Maximum Entropy Workshop, University of Wyoming.* 1981.

[5] H.A. Simon. Rational choice and the structure of the environment, *Physcal Review.* V. 63(2), 129-38, 1956.

[6] L.D. Stone. Theory of optimal search. *Operations Research Society of America, Arlington, Virginia.* ORSA Books, 1989.

Clinical Decision Support System for Diabetic Retinopathy Risk Evaluation

Emran Saleh [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
emran.saleh@estudiants.urv.cat

1 Introduction

Diabetic retinopathy (DR) is the main reason of a cumulative demolition of the retina for diabetic patients, being the essential cause of the vision loss among working-age adults. As diabetes prevalence grows, it does also the number of people suffering DR, being a main concern for health care centres. Frequent and early checking of eye fundus using non-mydriatic fundus cameras may minimize the risk of blindness development and the economic impact of the remedy as well [2]. Unfortunately, because of the large number of diabetic patients, it is not feasible to make a preventive screening to all the patients.

Physicians of the Ophthalmology Department of Sant Joan de Reus University Hospital (SJRU) did a statistical and clinical study which presented that 8% to 9% of diabetic patients developed DR [1]. Taking into account this low proportion, also supported by other studies, physicians concluded that some patients could be safely screened every 2 or 3 years and it would be better to focus the use of resources on those patients with more risk to develop DR.

The objective of this work to construct a *Clinical Decision Support System* (CDSS) to aid the clinicians to make a primary diagnosis and evaluate the risk of developing DR[3]. In this work we propose a CDSS based on an ensemble of decision trees called *Fuzzy Random Forest* (FRF). The goal of this model is to classify the new patients as healthy (no risk of DR) or sufferer (some sign of DR).

2 The data

The method proposed in the next section has been applied to data stored in the Electronic Health Records of patients which were methodically collected by the Ophthalmology Department of SJRU Hospital. The hospital provided

[*] PhD advisors: Antonio Moreno, Pedro Romero-Aroca, Aida Valls

us 2323 records of diabetic patients. The dataset consists of 579 patients with DR (class 1) and 1744 records of healthy patients (class 0).

To train and test the model we used the most relevant attributes, according to the previous study of the physicians [1]. Some attributes are numerical (e.g. Age and Body Mass Index), while the others are categorical (e.g. Sex). As the proposed model is based on fuzzy logic, we need to transform the input data into fuzzy sets. The numerical attributes are discretized into linguistic attributes. Each attribute is fuzzified into linguistic terms which are significative to the physicians.

3 Methods

The model proposed in this work is known as fuzzy random forest (FRF) which consists of a bunch of fuzzy decision trees (FDTs) as classifiers. Each branch of these classifiers is a rule that makes a decision. Various bootstrap samples of the training dataset have been utilized to assure the variety in constructing the fuzzy decision trees. To raise the diversity, random selection of a subset of the total attributes to split each node has also been used.

3.1 Random Forest Construction

The following are the essential steps to generate a random forest :

1. Pick random samples of the examples for training (bootstrap). The size of the bootstrap must be around 2/3 of the training dataset and the balanced distribution in the bootstrap must be taken into account.
2. Utilize each bootstrap to build a fuzzy decision tree (Section 3.2). To determine a new splitting of a tree node through the tree building process, a random subset of the attributes will be taken with size equal to γ.
3. Repeat steps 1, 2 until all fuzzy decision trees (n) have been built to comprise a random forest.

3.2 Fuzzy Decision Tree Induction

Many fuzzy decision trees induction methods have been proposed in the literature. The fuzzy decision tree induction algorithm proposed by Yuan and Shaw has been used in this work [4].

The steps of the induction process are the following:

1. Generate a subset of attributes of size γ, then select the best attribute for the root node v: the one with the **smallest ambiguity**.
2. Create a new branch for each of the values of the attribute v for which we have examples with support at least α.
3. Calculate the **truth level of classification** with a branch into each class.

4. If for at least one class, the truth level of classification is higher than β, terminate the branch with a label corresponding to the class with the highest truth level.
5. If the truth level is smaller than β for all the classes, check if there is a new node that can reduce the classification ambiguity.
6. If there is more than one attribute with lower classification ambiguity, select the attribute with the **smallest classification ambiguity with the accumulated evidence** as a new decision node from the branch. Repeat from step 2 until no further growth is possible.
7. If there are no attributes that can reduce the classification ambiguity, terminate the branch as a leaf with a label corresponding to the class with the highest truth level.

The parameters which used in the induction process are:

- The *significance level* (α) is used to filter if the evidence is relevant enough or not. If the membership degree of the evidence is lower than α , it is not used in the induction process.
- The *truth level threshold* (β) determines the minimum truth level of the conclusions obtained by the rules.

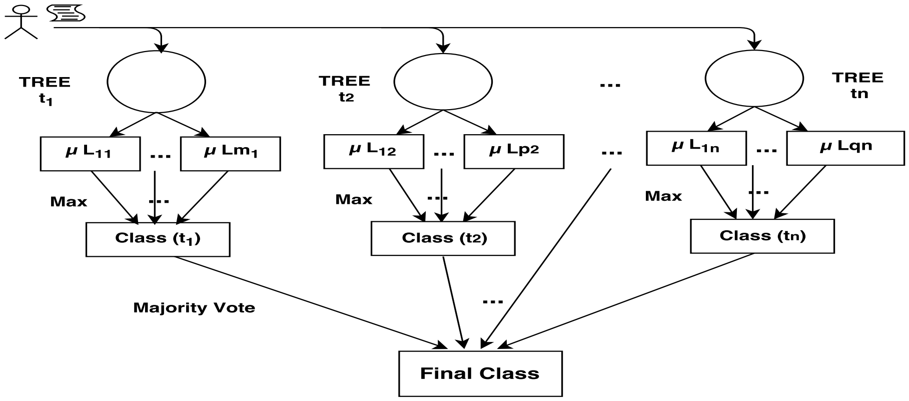


Fig. 1: Classification of a new observation in the Fuzzy Random Forest.

3.3 Random Forest classification

In a random forest, for an observation each rule of each tree gives a predicted class. Many techniques exist to obtain the final decision of the random forest. The following are the steps of the method used in this work to classify an observation (see Figure 1):

1. The new observation data is fed into each root node of the forest trees. Each branch (i.e. rule) of a tree gives an inference consisting of a class label of the leaf Lxy and a certain membership degree μ.
2. Aggregate all the inferences of the leaves of each tree to get one final decision of the tree (class label). The Mamdani inference procedure is utilized to predict the final class from each FDT: 1) Calculate the satisfaction degree of a rule using the t-norm minimum; 2) Calculate the membership to the conclusion class μ by multiplying the degree of support of the rule by the satisfaction degree; 3) Aggregate the memberships for the same class using the t-conorm maximum. To obtain a single final decision of a tree for an observation, we compare the highest membership degrees of the classes for this observation in order to check if the difference of the membership values is large enough to choose a final class. If the difference of the membership degrees is higher than a given threshold δ_1 then choose the class label of the highest membership value; otherwise, the class is "Unknown".
3. To obtain the final inference of the random forest, count the number of trees predicting each class. Choose the class label with the majority of votes if the difference between the two majority classes is higher or equal than a given threshold δ_2; otherwise, the class is "Unknown".

To ensure the observations are classified in a specific class with enough support, we defined two parameters, δ_1 and δ_2, in order to detect the cases where an observation belongs to different classes with similar memberships or a similar number of votes. In these cases, the observation is not classified (*i.e.* it is labelled as "Unknown").

Acknowledgement. This work was supported by the Spanish research projects PI15/01150 and PI12/01535 (Instituto de Salud Carlos III) and the URV grants 2014PFR-URV-B2-60 and 2015PFR-URV-B2-60.

References

[1] Pedro Romero-Aroca, Sofia de la Riva-Fernandez, Aida Valls-Mateu, Ramon Sagarra-Alamo, Antonio Moreno-Ribas, and Nuria Soler. Changes observed in diabetic retinopathy: eight-year follow-up of a spanish population. *British Journal of Ophthalmology*, page In press DOI, 2016.

[2] Pedro Romero Aroca, Javier Reyes Torres, Ramón Sagarra Alamo, José Basora Gallisa, Juan Fernández-Balart, Alicia Pareja Ríos, and Marc Baget-Bernaldiz. Resultados de la implantación de la cámara no midriática sobre la población diabética. *Salud cienc.*, 19(3):214–219, 2012.

[3] Emran Saleh, Aida Valls, Antonio Moreno, Pedro Romero-Aroca, Sofia de la Riva-Fernandez, and Ramon Sagarra-Alamo. Diabetic retinopathy risk estimation using fuzzy rules on electronic health record data. In *Modeling Decisions for Artificial Intelligence*. Springer, 2016.

[4] Yufei Yuan and Michael J Shaw. Induction of fuzzy decision trees. *Fuzzy Sets and systems*, 69(2):125–139, 1995.

Searchable Encryption for Geo-Referenced Data

Jordi Ribes González [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
jordi.ribes@urv.cat

Abstract. Searchable encryption schemes allow users to outsource a dataset in an encrypted form while preserving the ability to remotely and privately query over it. In this work we propose different techniques for searchable encryption that achieve range queries on two-dimensional geo-referenced data. The proposed techniques improve previous works from an efficiency and from a security point of view.

1 Introduction

The cloud computing paradigm offers very convenient data storage and computation services at a low cost, thus providing an attractive alternative to physical storage and self-managed servers. Nevertheless, even though cloud computing leads to many economical and functional benefits, the action of leaving data at the hands of an external cloud service provider poses many security and privacy concerns.

One way to address the security concerns that arise from the process of outsourcing data to the cloud is providing users with user-centered cryptographic techniques. However, it is not convenient to outsource encrypted data by using traditional encryption techniques, since any operation over the dataset must be carried out locally. To overcome this obstacle, alternative cryptographic schemes must be applied.

In recent years there have been important advances in cryptographic techniques that allow to take advantage of the cloud benefits while securing the data. For example, two of these techniques are homomorphic encryption and order-preserving encryption, allowing for remote computations and ordering on encrypted data respectively.

Searchable encryption [11,3,1,4,7] deals with the problem of remotely querying over encrypted data. By using searchable encryption schemes, it is possible to outsource a dataset in an encrypted form, while preserving the searching functionality by letting users be able to send encrypted queries to

[*] PhD advisor: Oriol Farràs Ventura

the cloud. In this way, users can remotely and securely query over encrypted data and retrieve the segment of the outsourced dataset satisfying the query conditions.

2 Searchable Encryption for Geo-Referenced Data

Our aim is to provide searchable encryption schemes that enable a client to delegate an encrypted version of a geo-referenced dataset to a semi-trusted, honest-but-curious server, in such a way that searching capabilities over the encrypted data are preserved.

In our setting, the client first delegates an encrypted version of its dataset to a server. Such a dataset consists of a collection of documents, each of which is attached to a particular geographical point. Afterwards, the same client may want to retrieve a subset of the outsourced dataset. By generating an encrypted query, it is able to recover the outsourced documents lying inside a chosen rectangular location.

Based mainly in the works by Shi et al. [10] and by Faber et al. [6], we develop four techniques for searchable encryption achieving two-dimensional range queries over encrypted data. These techniques show different efficiency and security trade-offs. The provided solutions are also general, in the sense that they make use of an arbitrary underlying keyword searchable encryption scheme. By changing this underlying scheme, different efficiency and security measures can be achieved.

We analyze the trade-off between performance, security and communication overhead of the presented options by considering the scheme by Cash et al [4] as the underlying searchable encryption scheme. Our solutions take advantage of the Boolean search and inverted index properties of [4].

As a novel approach with respect to previous works, we build on alternative combinatorial structures to lower the leakage of the schemes, thus improving security at the cost of increasing the query size and the search time. We also present a technique based on over-covers [6] that notably reduces the communication cost and the leakage of the queries at the expense of increasing the false-positive rate.

The proposed results have been presented at the 15th IFIP Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net 2016).

Acknowledgement. This work is partly supported by the European Commission under the H2020 project “CLARUS”, ref. H2020-ICT-2014-1-644024.

References

[1] D. Boneh, B .Waters. *Conjunctive, subset, and range queries on encrypted data.* In Proceedings of the 4th conference on Theory of cryptography (TCC'07), Salil

P. Vadhan (Ed.). Springer-Verlag, Berlin, Heidelberg, 535–554, 2007.

[2] C. Bösch, P. Hartel, W. Jonker, A. Peter. *A Survey of Provably Secure Searchable Encryption.* ACM Computing Surveys, 47 (2), 18:1–18:51, 2014.

[3] R. Curtmola, J. Garay, S. Kamara, R. Ostrovsky. *Searchable symmetric encryption: improved definitions and efficient constructions.* In Proceedings of the 13th ACM conference on Computer and communications security (CCS '06). ACM, New York, NY, USA, 79–88, 2006.

[4] D. Cash, S. Jarecki, C.S. Jutla, H. Krawczyk, M.-C. Rosu, M. Steiner. *Highly-Scalable Searchable Symmetric Encryption with Support for Boolean Queries.* CRYPTO, 353–373, 2013.

[5] *CLARUS: A framework for user centered privacy and security in the cloud.* Horizon 2020 project H2020-ICT-2014-1-644024. `http://www.clarussecure.eu/`.

[6] S. Faber, S. Jarecki, H. Krawczyk, Q. Nguyen, M.-C. Rosu, M. Steiner. *Rich Queries on Encrypted Data: Beyond Exact Matches.* ESORICS, 123–145, 2015.

[7] S. Kamara, C. Papamanthou, T. Roeder. *Dynamic searchable symmetric encryption.* In Proceedings of the 2012 ACM conference on Computer and communications security (CCS '12). ACM, New York, NY, USA, 965–976, 2012.

[8] J. Li, E. R. Omiecinski. *Efficiency and security trade-off in supporting range queries on encrypted databases.* In Proceedings of the 19th annual IFIP WG 11.3 working conference on Data and Applications Security. Springer-Verlag, Berlin, Heidelberg, 69–83, 2005.

[9] R. A. Popa, C. M. S. Redfield, N. Zeldovich, H. Balakrishnan. *CryptDB: Protecting confidentiality with encrypted query processing.* In Proceedings of the 23rd ACM Symposium on Operating Systems Principles (SOSP-11). 85–100, 2011.

[10] E. Shi, J. Bethencourt, T-H. Hubert Chan, D. Song, A. Perrig. *Multi-Dimensional Range Query over Encrypted Data.* In Proceedings of the 2007 IEEE Symposium on Security and Privacy. IEEE Computer Society, Washington, DC, USA, 350–364, 2007.

[11] D. X. Song, D. Wagner, A. Perrig. *Practical Techniques for Searches on Encrypted Data.* In Proceedings of the 2000 IEEE Symposium on Security and Privacy (SP '00). IEEE Computer Society, Washington, DC, USA, 44–, 2000.

On semi-automatic methods for localization and segmentation of optic disk in digital fundus images

José Escorcia-Gutierrez [⋆] [⋆⋆]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
joserafael.escorcia@estudiants.urv.cat

1 Introduction

Diabetic Retinopathy (DR) is an emerging blindness world epidemic due to its generative upgrowth of new blood vessels that nourish the retina. These vessels are responsible of an increment in the blood glucose level. As result, dilated small blood vessels (microaneurysms) and their rupture are the source of intra-retinal hemorrhages and fluid leaking composed by lipoproteins and lipids (exudates).

The contribution of this work is to present a comparative between Convexity Shape Prior and Grabcut algorithm for OD segmentation in color fundus images. Hence, the proposed algorithms are shown in Section 2. And finally, a brief Section 3 exposes the main conclusion.

2 Methodology

The proposed OD segmentation methods are based on the Discrete Convexity Shape [1] and Grabcut [5] algorithm. The general procedure is composed of three stages: (1) Preprocessing, which implicates the RGB green channel and the CIELAB lightness; (2) Segmentation of the main blood vessels located on the OD; (3) OD segmentation applying the two algorithms proposed. Finally, an accurate analysis of the results will be performed as well as a comparison. The next Figure 1 shows the described stages.

2.1 Preprocessing

An accurate OD segmentation needs to avoid false positives generated by the presence of blood vessels. Hence, we previously apply the *contrast-limited*

[⋆] PhD advisors: Domènec Puig, Aida Valls and Pedro Romero-Aroca

[⋆⋆] Collaborator: Jordina Torrents-Barrena

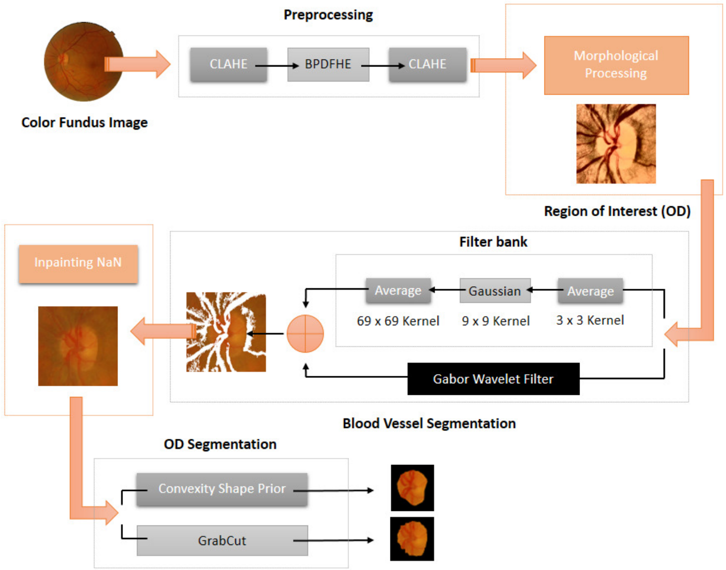


Fig. 1: Flow Chart of the proposed OD segmentation method.

adaptive histogram equalization (CLAHE) algorithm to compensate the non-uniform lighting effect. Then, in order to enhance blood vessels and make them more distinguishable, the *brightness preserving dynamic fuzzy histogram equalization* (BPDFHE) [4] method is applied on the previous response. Once we have the resulting image, CLAHE is finally executed another time.

2.2 Morphological processing

We propose a similar methodology as developed in [3] to obtain the region of interest where OD is located (following the steps outlined in Algorithm 1). In Figure 2 illustrates an example of fundus images that have been pre-processed through CLAHE in the presence of exudates. Note that, their appearance is similar to OD. Hence, it can be appreciated how bright structures (exudates) have been eliminated progressively until get the suitable region of interest.

2.3 Blood vessels segmentation

This section explains how the proposed bank filter composed by two Average filters and one Gaussian filter match with a Gabor Wavelet filter achieving effective detections. Concretely, the designed bank filter is supported in [2], where it is defined as a concatenation of the Average - Gaussian - Average sub-filters configured with three different kernel dimensions (*3 x 3*, *9 x 9* and *69 x 69*) and Gabor filters have been mostly used thanks to its performance as feature extractor. In order to detect the whole structure, Gabor Wavelet filter has been generated from a mother wavelet and configured with an arrangement of *8 orientations* and *5 scales*.

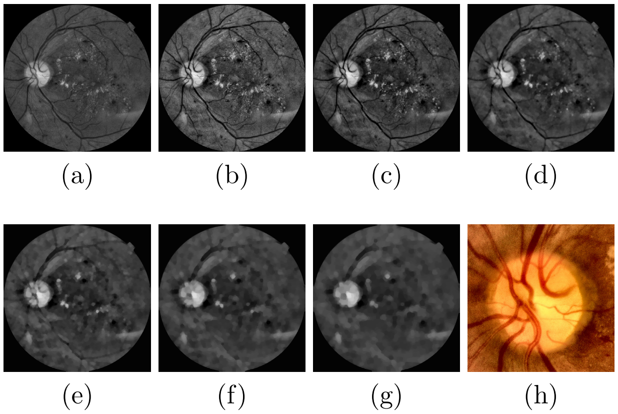


Fig. 2: Morphological procedure to obtain the OD region. (a) G channel. (b) Preprocessed image through CLAHE algorithm. (c) to (g) Morphological opening and closing operations with radius equivalent to 4, 8, 12, 16 and 20 pixels, respectively. At the end, (h) represents the obtained OD region of interest.

Algorithm 1 Morphological procedure to detect OD region

```
Require: $G_{CLAHE}$
Ensure: $OD_{Detection}$
$StructuralElement \leftarrow Disc$
$Radius \leftarrow 4$
$G_{\text{closing}} \leftarrow G_{CLAHE}$
while $OD_{Detection} = true$ do
    $G_{\text{opening}} \leftarrow imopen(G_{closing}, StructuralElement, Radius)$
    $G_{\text{closing}} \leftarrow imclose(G_{opening}, StructuralElement, Radius)$
    $G_{\text{BW}} \leftarrow Threshold_{\text{Otsu}}$ {Number of pixels determined the $OD_{Detection}$ is true or false}
    if $OD_{Detection} = \text{true}$ then
        return true
    end if
    $Radius \leftarrow Radius + 4$ {Increasing factor by 4}
end while
```

2.4 Optic disk segmentation

Regarding this phase, two different algorithms are proposed to accomplish OD segmentation. Both of them are effective combinatorial optimization techniques based on prior information such as shape, color separation and geometric interactions. Firstly, we adapt the new Convexity Shape Prior algorithm [1] to be used on medical image applications. Next, the obtained results are compared with the traditional and iterative Grabcut approach [5]. In Figure 3 shows a set of segmentation results validated according to the doctor's experience.

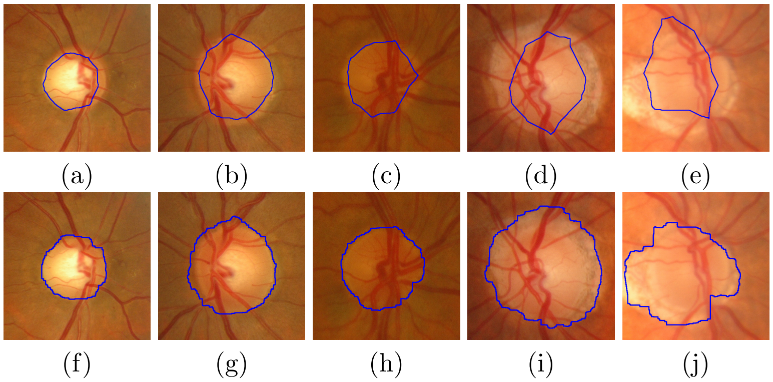


Fig. 3: OD segmentation result illustrated by blue boundaries. The first row contains the results of Convexity Shape Prior algorithm and the second row corresponds to Grabcut algorithm.

3 Conclusions

An analysis of two algorithms (Convexity Shape Prior and GrabCut) are introduced for OD interactive segmentation. During the preprocessing stage, blood vessels are enhanced with both CLAHE and BPDFHE methods to improve their contrast. Next, in order to eliminate the presence of blood vessels inside the OD structure, a new matching filter has been designed. In addition, it is possible to increase the OD accuracy by interpolating the statistical color information of the neighbors using an inpaint NaN algorithm. At the end, the OD segmentation algorithms are applied without any dependence related to a predefined shape.

References

[1] Lena Gorelick, Olga Veksler, Yuri Boykov, and Claudia Nieuwenhuis. Convexity shape prior for segmentation. In *Computer Vision–ECCV 2014*, pages 675–690. Springer, 2014.

[2] Diego Marín, Arturo Aquino, Manuel Emilio Gegúndez-Arias, and José Manuel Bravo. A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. *Medical Imaging, IEEE Transactions on*, 30(1):146–158, 2011.

[3] Diego Marin, Manuel E Gegundez-Arias, Angel Suero, and Jose M Bravo. Obtaining optic disc center and pixel region by automatic thresholding methods on morphologically processed fundus images. *Computer methods and programs in biomedicine*, 118(2):173–185, 2015.

[4] Fatemeh Rezazadeh, Jamshid Shanbehzadeh, and Hossein Sarrafzadeh. Brightness preserving fuzzy dynamic histogram equalization. 2013.

[5] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

Classification of computer technologies for multimorbidity

Wilfrido Ortega León *

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
wilfrido.ortega@estudiants.urv.cat

Abstract. Clinical guidelines are valuable instruments to record and transmit available evidence based knowledge. Several medical informatics technologies have been developed to merge computer knowledge representation relative to diseases active in a multimorbidity. We propose a classification of the current available technologies and assess their strengths and weaknesses.

Keywords: multimorbidity, co-morbidity, clinical guidelines, knowledge acquisition

1 Background

Progress in healthcare and medical treatment have contributed to a large segment of the population living longer with associated chronic conditions [1]. This phenomenon is associated to medical conditions involving co-morbidity and multimorbidity.

2 Introduction

Multimorbidity is the coexistence of multiple long term diseases in the same individual at the same time, with none of the diseases being more prominent than the others [2,1] as figure 1 depicts. Multimorbidity doesn't have well-defined criteria for medical diagnosis [1]. The presence of multimorbidity leads to uncovering others problems closely related to multimorbidity such as disability (difficulty of lack of independence), frailty (state of vulnerability) and patient complexity (dealing with medical, social and behavioral factors) [1].

Co-morbidity, on the other hand, is defined as the coexistence of secondary diseases associated to a primary disease or index, as figure 2 shows. Treatment is driven by the primary disease, but it is adapted to include treatment for the secondary diseases [2,3].

* PhD advisor: David Riaño Ramos

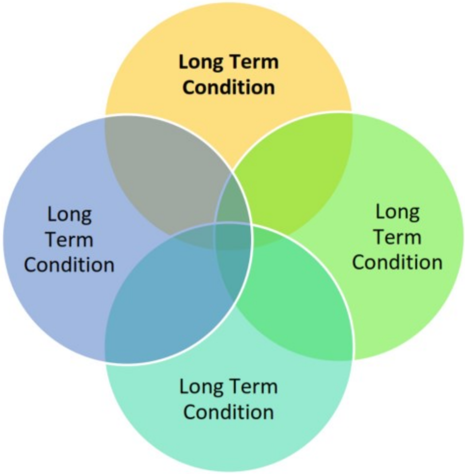


Fig. 1: Diagram of multimorbidity.

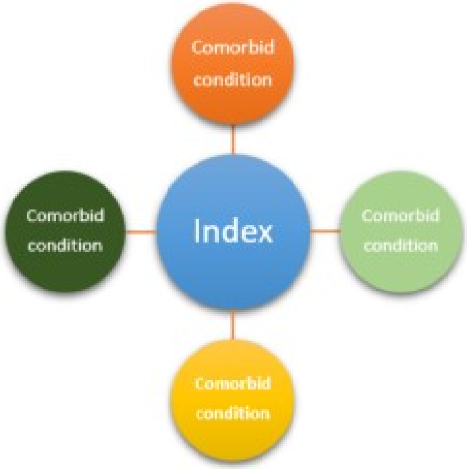


Fig. 2: Diagram of co-morbidity.

Formal languages are suitable instruments for capturing knowledge embedded in clinical practice guidelines [1]. Thus, from a formal language perspective knowledge acquisition and machine learning are suitable instruments for capturing knowledge contained in clinical practice guidelines. Medical practitioners utilize computers that use computer structures capable of capturing medical knowledge contained in clinical practical guidelines [4].

Specifically, the knowledge acquisition approach evolves from clinical experts interpreting clinical guidelines into computer structures done by knowledge engineers [4,5]. Knowledge acquisition is characterized by being evidence-based, but time consuming to implement.

The machine learning approach uses the information in databases related to the management of multimorbid patients. Then, it generalizes past individual experiences by constructing computer structures [6,7,8,9]. Developing and building computer structures for managing multimorbid patients is a promissory new field of active research [10].

According to Abidi et al [11] and Jafarpour [12] the combination of knowledge in the knowledge acquisition approach happens at specific places along the path of a transformation process that begins at clinical guidelines and goes all the way to the end as computer structures.

Abidi [13] suggested an additional proposal claiming that knowledge of the diseases of multimorbid patients can be combined at modeling level and at the execution level. The former is related to computer structures generated for each disease affecting multimorbid cases are combined into a single computer structure ready to be use by medical practitioners on multimorbid cases. The later takes place when each one of the computer structures of the diseases of multimorbid cases are executed and their results are integrated into a single guideline.

Jafarpour extended the modelling and execution level approach by adding two additional points of combinations. The first point is at the guideline level and the second point is at the computerization level [12].

3 Objectives

To perform an extensive literature review of technologies for computer-based health-care combination to manage multimorbid patients. Details on the technologies and their analytical comparison related to their strength, weaknesses and maturity will be explored.

4 Classification of Methods

Knowledge integration. The knowledge integration method (KIM) is a general method of classification that includes ontology merging, logic and constraint satisfaction, transition fitting. All of these approaches have in common the explicit representation of conflicts experienced by single-disease guidelines.

Treatment Integration method (TIM) involves drug integration checking that relates the pharmacy knowledge on treatments to detect drug interactions (side effects) [14,15,16,17]. Another approach is the drug interaction resolution [18] that involves the implementation of rules aimed at solving conflicts upon detection and suggest physicians a possible conflict-free treatment.

Clinical pathway pattern discovery (CPPD) uses clinical logs to infer pathways patterns resulting from joint probabilistic models.

State decision-action induction (SDAI) extracts clinical algorithms from episodes of care. CPM, CPPD and SDAI have in common that knowledge is extracted from past clinical interventions.

Data integration approach deals with analysis of episodes of care of multiple patients to discover hidden patterns. Clinical process mining (CPM) analyzes clinical logs on sequential treatments on patients. The analysis of these logs result in multimorbid treatment models after the identification of common sequences, concurrences and branching of clinical actions mined in the process. Additional technologies in this area are dynamic programming optimization and latent Dirichlet allocation (LDA) combined with Gibbs sampling [19]. Riaño et al [20] proposed a four-step process to obtain clinical algorithms that generalize treatments described in sets of episodes of care.

5 Discussions

Currently, clinical guidelines are the best source for dealing with diagnosis, prognosis and treatment. The information in clinical guidelines is very specific and comes from controlled situations/experiments. Therefore, it is rich in internal validity, but low in external validity. In real world cases, controlled situations do not occur due to heterogeneity of multimorbid patients and high cost of episodes of care performed on controlled condition settings.

References

[1] Fortin M, Mercer S, Salisbury C. Introduction to Multimorbidity. In: ABC of Multimorbidity. Wiley 2014. 1-4

[2] Boyd CM, Fortin M. Future of multimorbidity research: how should understanding of multimorbidity inform health system design? Public Health Reviews 2010;32(2):451-474.

[3] Jakovljević M, Ostojić L. Comorbidity and multimorbidity in medicine today: challenges and opportunities for bringing separated branches of medicine closer to each other. Psychiatr Danub. 2013 Jun;25 Suppl 1:18-28.

[4] Peleg M. Computer-interpretable clinical guidelines: a methodological review. J Biomed Inform. 2013 Aug;46(4):744-63.

[5] de Clercq PA, Blom JA, Korsten HH, Hasman A. Approaches for creating computer-interpretable guidelines that facilitate decision support. Artif Intell Med. 2004 May;31(1):1-27.

[6] Bohada JA, Riaño D, , López-Vallverdú JA. Automatic generation of clinical algorithms within the state-decision-action model. Expert Syst Appl. 2012;39(12):10709–10721.

[7] Maruster L, van der Aalst W, Weijters T, van den Bosch A. Automated Discovery of Workflow Models from Hospital Data. Proc. Dutch-Belgian Art Int Conf. 2001:183-190.

[8] Mans R, Schonenberg H, Leonardi G, Panzarasa S, Cavallini A, Quaglini S, van der Aalst W. Process mining techniques: an application to stroke care. Stud Health Technol Inform. 2008;136:573-8.

[9] Mans RS, Schonenberg MH, Song M, van der Aalst WMP, Bakker PJM, Daelemans W. Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital. In Biomedical Engineering Systems and Technologies 25:425-438.

[10] Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, Campbell E, Bates DW. Grand challenges in clinical decision support. J Biomed Inform. 2008 Apr;41(2):387-92.

[11] Abidi SR, Abidi SSR. Towards the merging of multiple clinical protocols and guidelines via ontology-driven modeling. Proc. Artificial Intelligence in Medicine 2009;5451:81-85.

[12] Jafarpour B. Ontology merging using semantically-defined merge criteria and OWL reasoning services: towards execution-time merging of multiple clinical workflows To handle comorbidities. PhD Thesis. 2014.

[13] Abidi SR. A conceptual framework for ontology based automating and merging of clinical pathways of comorbidities. 2009;LNCS 5626:55-66.

[14] Epocrates drug interaction checker. Available from URL:http://online.epocrates.com/interaction-check. Accessed Feb16, 2016.

[15] LexicompR drug interaction checker. Wolters Kluwer. Available from URL: http://www.wolterskluwercdi.com/lexicomp-online/. Accessed Feb 16, 2016.

[16] WebMD drug interaction checker. Available from URL: http://www.webmd.com/interaction-checker/. Accessed Feb 16, 2016.

[17] Grando A, Farrish S, Boyd C, Boxwala A. Ontological approach for safe and effective polypharmacy prescription. AMIA Annu Symp Proc 2012;2012:291-300.

[18] López-Vallverdú JA, Riaño D, Collado A. Rule-based combination of comorbid treatments for chronic diseases applied to hypertension, diabetes mellitus and heart failure. In: Process Support and Knowledge Representation in Health Care. Springer 2013. 30-41.

[19] Newman D, Asuncion A, Smyth P, Welling M. Distributed algorithms for topic models. J Mach Learn Res 2009;10:1801–1828.

[20] Riaño D, Collado A. Model-based combination of treatments for the management of chronic comorbid patients. AIME 2013. LNCS 7885: 11-16.

This proceeding book contains the contributions presented at the 3rd URV Doctoral workshop in Computer Science and Mathematics. The main aim of this workshop is to promote the dissemination of the ideas, methods and results that are developed by the students of our PhD program.

[DΣIM] Departament d'Enginyeria Informàtica i Matemàtiques

ETSE Escola Tècnica Superior d'Enginyeria
UNIVERSITAT ROVIRA I VIRGILI